

Transatlantic tax dilemmas

The British government has too much tax revenue for comfort, and the US government too little. The former needs reform, the latter needs higher taxes.

THE US and British governments are in curiously different dilemmas about their budgets for the year ahead. The federal government, still wrestling with its huge deficit, is having to cut back on public spending dear to at least the administration's heart (see p. 646); last week's budget for the US financial year beginning on 1 October will nevertheless make only a small difference to the rate at which public indebtedness accumulates, which means further troubles ahead. In Britain, by contrast, the government suffers from a literal embarrassment of riches. It now seems certain that it will be in surplus for the present financial year when that finishes at the end of March. For the year ahead, the prospects are even brighter — economists and stockbrokers are talking of a surplus exceeding £6,000 million if nothing is done to change the pattern of tax rates and public spending in the budget to be published on 18 March. That is why the British Chancellor of the Exchequer, Mr Nigel Lawson, is being deluged with advice about the better balancing of his books. How should one "spend" £6,000 million or more?

Lawson has less room for manoeuvre than he would like, for reasons well documented by the painful experiences of recent decades. The British economy appears unable to sustain rapid economic growth without running into trouble — domestically, increased inflation; externally, a deteriorating balance between exports and imports. The causes of this phenomenon, not properly understood, lie in the pattern of industrial production and the inflexibility of even successful industries in supplying increasing demand for their products. After two or three years of respectable growth, Britain's external balance has indeed slipped into the red, and while inflation last month (at 3.5 per cent) was a record low, the months ahead may see a deterioration.

This is why the chancellor cannot simply give away his prospective surplus without risking trouble. The same applies to the popular cause of using the surplus to make the National Health Service function properly again (see *Nature* 330, 590; 1987) for that would have the effect of transferring the budget surplus into salaries. The same objection would no doubt be made to the proposal to use the prospective surplus to make good the damage done to research and higher education over the past ten years and more, but that is a short-sighted view. The truth is that the British government does not yet appreciate how much damage has been done — and that one consequence will be shortages of skill likely further to impede industry's capacity to respond flexibly to changing circumstances.

Reform

Yet Lawson cannot just keep the money, and is indeed committed to the cause of tax reform, partly to reduce the high rates of tax on higher incomes (the maximum is 60 per cent) and partly to remove anomalies in the present system which provide people with low incomes with various disincentives to work. Unfortunately, the general British opinion that next month's budget is, against the prospect of the surpluses to come, an historic opportunity for change must be moderated by the way past political promises have further constrained his freedom.

The US tax reform of 1986 is a model that Lawson cannot follow. The breathtaking simplicity of that device for creating a

single rate of tax on all income (or two rates for an interim period) was purchased by the abolition of a host of tax shelters that the British chancellor has promised to leave intact — tax relief on mortgage payments and contributions towards pension funds, but not on other kinds of personal savings, for example. Is it any wonder that successive governments complain that British savers invest too little in productive industry? Similarly, the British government, for no good reason, declines to extend its tax on spending (called value-added tax) from the general run of personal spending to, for example, food. The justification for perpetuating in this way an economic distortion of the pattern of people's spending is that extending value-added tax to all purchases would penalize the poor (but the issue has also become part of the government's determination not to be thought too compliant to rulings in Brussels). At some stage, there may be a government brave enough to bite these bullets.

Logic

Meanwhile, in Britain, the only logical route to tax reform is that advocated in the early 1970s by the Conservative government of Mr Edward Heath — linking the income-tax system with that for paying benefits to the unemployed or those whose salaries are too low for anybody's comfort. That is the direction in which the government should make a start next month. Some among its successors would naturally use an integrated system of this kind to redress the injustices that are accumulating in the present systems of benefits, but that will happen anyway.

Curiously, even by the government's own lights, there is also a strong case for creating some tax shelters that do not now exist in Britain, where people's private spending on health and education must be made from income after tax. Can that continue to make sense when the government is urging both customers and the institutions providing them with these services to rely less on government funds and more on funds from private sources? What chance is there that more private persons will kit themselves out with health insurance, as the present government hopes, when they must pay the cost from taxed income? Why should young people seeking extra skills do so readily when fees for the courses they would follow are similarly payable after tax? And why should private people behave generously towards even the educational institutions they admire when they must first pay taxes on their gifts? Not all these stratagems would be popular with all the present government's putative successors, but Lawson cannot logically fail to follow them.

The position in the United States is arithmetically simpler if politically just as complicated. The budget deficit amounts to 15 per cent of government spending, but nobody pretends that it should be made to vanish overnight. The modified Gramm-Rudman Act would make it shrink to zero in the three years ahead, but the administration has evidently been counting on economic growth (and the disproportionately higher tax revenues that follow) to pull most of its chestnuts out of the fire. But one of the benefits of a simplified tax system is that it can be simply changed. If the standard rate of tax were, for example, increased from 28 per cent to 29 per cent, much of its problem would be solved. Is that not worth a try? □



An unfound problem

For more than two years, Japan has been working on an imaginative idea. It needs more help.

JAPAN deserves a more generous response to the plan for a programme of research called the Human Frontier Science Program than it has been given so far. So much should by now be plain. When the project was first launched two years ago, with the backing of the then Prime Minister Mr Yasuhiro Nakasone, the intention was to kill several — probably too many — birds with a single stone. One objective was to deflect pressure from the United States over the continuing trade imbalance by demonstrating Japan's willingness to spend large sums of money (the equivalent in yen of \$6,000 million was mentioned at the outset) altruistically on an internationally coordinated programme of research. Another was to demonstrate Japanese competence in some field of endeavour not driven by commerce. Yet another was to make a mark. It is not unreasonable also to guess that part of the hidden agenda of the Human Frontier Science Program is the wish of basic researchers and their backers in Japan, not least the Science and Technology Agency and even the Ministry of International Trade and Industry, to find a way to persuade the powerful Ministry of Finance to do what it should have been prepared to do years ago — modify its open scepticism of the value of academic research.

The weakness in the case for the Human Frontier Science Program is that, from the outset, it has been a programme in search of a project. On several occasions and at several different levels, Japanese officials and researchers have taken to asking their putative partners to suggest projects that might with advantage become central to the Human Frontier Science Program. Most of the answers delivered have been unsatisfying, partly because the question has been too vaguely put, partly because the hidden agenda has not been explicit and partly because the respondents, mostly people whose daily domestic battle is that of winning a little extra to support their own projects, do not easily rise to the extent of Japan's challenge. Too often, governments have sent diplomats or administrators to meetings at which scientists would have been more appropriate. Suggestions that Japan should take on special responsibility for certain problems of the developing world have made little impact, when the hidden agenda is that of finding a formula for supporting first-world science.

There is no easy way of telling how the Japanese dilemma is to be resolved, but governments elsewhere, and that of the United States in particular, should respond to it more imaginatively than they have done so far. The principle was endorsed by the governments participating at the economic summit at Venice last summer, but there has not yet been an official response from elsewhere. Meanwhile, Japan is into the second year of what is called a "feasibility study" which, to judge from the document published in Tokyo last month, is a device for making lists of the important fields of biological research.

There are several fields of research at once important and interdisciplinary enough to be a substrate for Japan's ambitions. If the project to sequence the human genome were regarded not just as a once-and-for-all task (which is silly, in any case), but as the core of a rounded programme of work in human genetics with implications for medicine on the one hand and the physiology of the gene on the other, that would suffice. More to Japan's taste might be the kind of programme taking the analogy between neural networks and the physiology of the brain as its core (see p. 657) as its starting point. Perhaps the programme should include both. What would be required of partner governments is that they should be in a position to know what work is being supported in the chosen fields at their own institutions, that there should be systematic ways of exchanging information and that there should be a general understanding on coordination, especially in the training of people. What the government

of Japan should do, without more beating about the bush, is to appoint a group of able people from elsewhere, acting in their personal capacities and not as official representatives, who could both hammer out the specification of a programme saleable to governments and devise a system for managing it that would be acceptable elsewhere. □

Graduate submission

A study of British graduate education casts doubt on politicians' understanding of universities.

READERS who are variously distressed and bored by the continuing saga of the British universities should take care; many of the forces that have helped to impoverish British institutions are also gathering strength elsewhere. Moreover, the subtlety of the process by which the autonomy of academic institutions is threatened by their dependence on public funds requires constant vigilance, which is nicely illustrated by a document published last week by the British parliament's Public Accounts Committee (PAC). This goes a long way to offset the welcome that will be given to the Secretary of State for Education and Science if indeed he delivers on his promise last week (see page 645) to moderate the oppressive measures in his education bill. For the PAC document displays a profound ignorance of what universities do by people who should know better.

The circumstances are these. Most graduate students at British universities are supported financially by the government, which provides the funds enabling the research councils to pay stipends to students chosen to follow graduate courses by universities. The big provider is the Science and Engineering Research Council (SERC), but the other four councils deal with research students in their fields as does the British Academy in the humanities. The stipends are tax-free but inadequate; there is also an inadequate allowance for "research expenses" for the university department concerned. Those who count candle-ends have been muttering about the system for the past decade, chiefly after the discovery that many of those holding post-graduate awards in the social sciences apparently never got round to completing a thesis, submitting it and being examined on its content. Under pressure, the Economic and Social Research Council has taken to blacklisting departments judged inadequate by this simple indicator.

The twin themes of the PAC report are what is called the "submission rate" (the proportion of graduate students submitting theses within four years) and the doctrine that graduate study is primarily meant to "meet the nation's future manpower needs". Witnesses before the committee, promised that they would do everything in their power to meet the committee's goal of a submission rate "well beyond . . . 50 per cent". PAC also asks for central monitoring of the awards system, partly in the hope against hope that better forecasting of "manpower needs" would then be possible.

The errors in the argument are plain. Students who fail to submit a thesis rarely do so from idleness, but are more likely to have nothing to put into it. Institutions seeking to meet PAC's goal have a simple remedy — make research projects simpler. But that will mostly be at odds with their concepts of academic standards. Should they behave differently because a powerful committee says so? The argument on manpower planning is similarly faulty, at least on the assumption that British graduate education is entirely in the gift of the research councils, which overlooks the growing (but belated) trend for industrial companies to safeguard their own needs for trained people. But assuming PAC's implicit justification of its inquiry is that public money is spent on stipends to graduate students, it would do better to ask whether research councils should continue to award quotas to university departments (where selection criteria are far from uniform) and whether these costs are best met from a research budget already seriously strained. □

UK government concedes to critics of education bill

- Baker backs down after fierce attacks
- British academics' lobbying pays off

London

FACED with mounting hostility from its own supporters, the British government last week waved the white flag and pledged to amend its proposals on the financing and management of higher education, set down in the Education Reform Bill. Objections have centred around clauses in the bill that would empower the Secretary of State for Education and Science to interfere with the administration of individual institutions. The government had clearly not anticipated the ferocity of the attacks on the bill from the university sector and its own backbench Members of Parliament. University vice-chancellors, whose orchestrated campaign of lobbying seems to have paid off (in spite of its slowness to get off the ground), are reserving final judgement on the government's new stance until specific details of the amendments are tabled.

Announcing the government's climb-down, Mr Kenneth Baker, Secretary of State for Education and Science, told the House of Commons standing committee that is considering the bill that much of the criticism had been "both misguided and misleading, imputing to the government all manner of sinister motives which it simply does not have".

The bill proposes to replace the existing University Grants Committee (UGC) with a Universities Funding Council (a similar 'Polytechnics and Colleges Funding Council' will also be set up), which would be able to attach conditions to any payment it made to an institution. Any institution that failed to comply with the conditions could be required to repay any sums received from the council, with interest. Furthermore, the bill would enable the secretary of state to attach strings to any grants he made to the funding councils. The bill does not make clear whether the funding councils' powers of direction over institutions would apply to sums received from non-government sources.

Baker has promised to bring forward amendments at a later stage of the bill's passage through parliament that will:

- make explicit that funding councils could attach conditions only to funds received from the councils;
- make clear that the secretary of state would not be able to impose specific conditions in relation to the flow of money to particular institutions;

- make any direction to the funding councils by the secretary of state subject to negative resolution by both houses of parliament;
- empower funding councils to give advice to the government;
- qualify the secretary of state's power to confer additional functions on the funding councils;
- clarify the power of the funding councils to seek repayment of funds not used in accordance with a particular condition.



Baker: bill will allow "greater independence".

Professor Sir Mark Richmond, chairman of the Committee of Vice-Chancellors and Principals, cautiously welcomed Baker's announcement as a "step in the right direction". The Association of University Teachers (AUT) was similarly thankful for the government's "surprise concessions". Throughout the controversy over the offending clauses, Baker has insisted that the secretary of state's powers of direction over the funding councils would be used only as a last resort, and that theoretically the education secretary at present has unlimited powers over the UGC. "The only change is that it is now set down in black and white for all to see", he told the standing committee last week.

The education spokesman for the Liberal Party, Mr Paddy Ashdown, contended that legislation spelling out the education secretary's powers would erode the powerful convention that at present effectively forbids him to direct the UGC. In response to government claims that statutory powers of direction under the Science and Technology Act of 1965 exist in relation to the research councils but

Livermore lab head

JOHN Nuckolls has been named as the new director of Lawrence Livermore National Laboratory (LLNL), to succeed Roger Batzel, who will step down on 3 April after directing the laboratory for more than 16 years.

Nuckolls' selection is seen by some as a victory for supporters of the Strategic Defense Initiative (SDI). Physicist Edward Teller, a strong proponent of SDI, is said to have lobbied hard for his selection.

Although he was Teller's choice, Nuckolls is expected to be a stronger leader than Batzel, leaving individuals like Teller little opportunity to express their personal views as those of the laboratory. Nuckolls is known to be a strong supporter of new weapons technology, but his colleagues nevertheless expect him to be objective and realistic in his role as scientific adviser to Washington policy-makers.

Sidney Drell, co-director of the Center for International Security and Arms Control at Stanford University, praises Nuckolls as "a fine scientist", whose scientific perspective and intelligence should make him a balanced leader. "I would not label him as an ideologue", said Drell.

M.B.

TV satellite failure

THE West German Minister for Telecommunications said in Bonn this week that there "is almost no hope" of saving the TV-SAT direct-broadcasting satellite plagued by a failed solar panel (see *Nature* 331, 200; 1988). The satellite would have marked Europe's entry into direct satellite television broadcasting. Its failure is a bitter disappointment to the German Bundespost, exacerbated by the jubilation surrounding its launch on the Ariane rocket that confirmed resumption of Europe's commercial space industry last November. If West Germany does decide to abandon efforts to save the satellite, only DM95 million of its DM390 million cost will be recovered from the insurers. P.C.

have never been used, Ashdown quoted from a letter written by Sir John Kingman, a former chairman of the Science and Engineering Research Council (SERC), to Sir David Hancock, permanent secretary at the Department of Education and Science (DES).

Although conceding that during his term of office he never received formal direction, Kingman says he did receive "frequent instruction and requests from DES officials. Had I not complied with these I would have been in breach of the act or of the SERC charter, or (as I was often reminded) liable to receive formal direction under the act. The power of direction, like the nuclear deterrent, is no less effective for remaining unused."

Simon Hadlington

US budget does well by big science in 1989 proposal to Congress

Washington

PRESIDENT Reagan's \$1.094 billion 1989 budget, released last week, does well by big science. The Superconducting Super Collider (SSC), the space station and the Strategic Defense Initiative (SDI) are all in line for healthy funding. But Congress may not be willing to join the White House in committing the United States to large projects for which the major costs will be felt only after President Reagan has left office. Even in less contentious areas, the budget that ultimately emerges from Congress may be quite different from the blueprint the White House has proposed.

The proposed budget represents a compromise between philosophical ideals and political realities. Restrained by an agreement worked out with Congress last November to reduce the federal deficit, the president's swansong budget is more muted in its support for ideologically contentious programmes.

The massive military buildup begun by the Reagan administration slows considerably, although prized projects such as SDI retain strong support. Much of the rationale for federal spending on science and technology remains tied to strengthening economic competitiveness. The

budget calls for spending on non-military basic research to rise some 6 per cent to \$10,300 million for 1989.

National Aeronautics and Space Administration (NASA)

Space activities receive a significant influx of new money from the president's budget. The increase is related to the planned resumption of shuttle flights, now scheduled for this summer, the first construction phase for the international permanently manned space station.

Only one new space science mission is included in NASA's budget request; the Advanced X-Ray Facility, one of the so-called "great observatories" planned to accompany the Hubble Space Telescope and the planned Gamma-Ray Observatory. The comet rendezvous, asteroid fly-by mission (CRAF), is passed over, although missions to Mars (Mars-Observatory), Venus (Magellan) and Jupiter (Galileo) continue to receive support. The administration is seeking \$100 million for the Pathfinder Program, a new advanced technology initiative.

National Science Foundation (NSF)

The plans announced last year to double the NSF budget over the next 5 years got off to a shaky start in 1988, with Congress appropriating funds amounting to only a 6 per cent increase over 1987. But director Erich Bloch says the administration remains committed to its 5-year plan, and the new budget request reflects that. Support for the proposed multidisciplinary Science and Technology Centers appears as a separate line item for the first time. The \$150-million request would cover the full 5-year cost of 12–15 centres. NSF hopes that such 'up-front' funding would protect the centres from future budgetary fluctuations.

The administration is also seeking big increases for NSF supercomputing activities. A 53-per-cent increase in money for networking and communications research and infrastructure will permit work to continue on NSFNET, NSF's high-speed computer network.

Health and Human Services

After the administration's plan last year to cut the budget of the National Institutes of Health (NIH) by \$750 million was thwarted by Congress, its request this year rises 5.4 per cent. Most of this increase is requested for research project grants.

The number of new grants (5,611) is fewer than in 1988 but is enough to bring the total to a record 20,600. The amount available for each grant would increase by about 3 per cent for new grants and just under 5 per cent for existing ones.

Intramural research is treated less generously, budgeted with barely enough to

Polish physicist gets a job at last

London

PROFESSOR Andrzej Jurewicz, one of the pro-Solidarity physicists from Warsaw's former Institute of Nuclear Research, who lost his job when the institute was reorganized into three new establishments at the end of 1982, has at last found work after almost five years. The reorganization of the institute made it possible to purge it of political "undesirables" without contravening the labour laws on academic tenure. Jurewicz's new post, which carries the rank of *docent* (associate professor) is at the Space Research Centre in Warsaw.

Commenting on Jurewicz's new appointment, *W Okopach* (In the trenches), the underground Solidarity bulletin of the Warsaw nuclear research establishments (No. 93, 20 January 1988), said: "No one ever told him, right up to the end, why he had been thrown out. We believe that a certain role in the resolution of this kafkaesque situation was played by an article in *Nature* (No. 6121), which created a considerable stir". Vera Rich

cover inflation (up 3.7 per cent on 1988). Support for the human genome project has been increased from \$17 million to \$27 million, with the emphasis on identifying chromosomal markers close to genes associated with known genetic disorders.

Research into AIDS retains its high priority and is now dealt with as a separate budget item. Of the \$1,300 million requested, \$400 million is to be spent on education. NIH would receive \$588 million, \$120 million more than they will spend on AIDS research in 1988. The Centers for Disease Control (CDC), responsible for a nationwide public information campaign, would receive \$401 million compared with \$301 million in 1988. Although CDC's AIDS activities will increase, their other programmes do not keep pace with inflation in the proposed budget.

Department of Energy

Of the \$16,080 million requested by the Department of Energy (DoE), slightly less than half goes to defence activities. But Under Secretary Joseph Salgado, announcing the numbers, emphasized that the proportion of DoE spending devoted to defence has fallen from the 68 per cent spent two years ago, with a number of big civil science projects, notably the SSC, as the beneficiaries.

The \$363 million requested to begin construction of SSC is essentially the same amount as was requested and turned down by Congress last year, when only \$25 million for continuing research and development was made available. And although DoE also hopes to start building a 6-GeV synchrotron light source at

1989 budget request (millions of dollars)

Agency	1989 Request	% Change from 1988
DoD	299,529	2.8
Research	38,800	2.0
SDI	4,950*	39.8
NSF	2,050	19.4
Research	1,603	10.3
Education	156	12.2
Antarctic pgms	141	12.8
S & T centres	150	N/A
NASA	11,488	27.3
Space station	967	46.7
Space science	1,860	18.0
DoE	16,080	13.9
Research	1,713	25.3
SSC	363	1,352.0
Human genome	19	76.6
Superconductivity	95	41.2
HHS	464,800	8.6
NIH	6,535	5.4
CDC	523	1.4
AIDS	1,300	36.7
Human genome	27	58.8
Interior	5,950	-9.6
USGS	425	-5.1
Commerce	2,423	-0.6
NOAA	1,121	-2.9
EPA	3,129	-24.1

* Includes \$400 million from DoE.

Argonne National Laboratory, the 'big science' community does not escape scot-free: about \$60 million has been deleted from the basic energy science research budget, which includes much support of high-energy physics experimentation.

Research into superconductivity, covering traditional 'low-temperature' materials as well as the new superconductors, benefits to the tune of \$95 million, and a project to collaborate with industry on the exploitation of high-temperature superconductors is planned for Los Alamos National Laboratory.

A sum of \$19 million has been set aside for work on the human genome project, in collaboration with NIH.

Department of Defense

SDI retains high priority with a 40 per cent increase requested, well above the 2.8 per cent rise scheduled for total defence spending, and the 2.0 per cent for non-SDI research programmes. But the projected \$4,950-million expenditure is still almost \$2,000 million below that planned by the administration a year ago and will require cuts in some SDI projects.

The trend towards research on high velocity interceptor rockets and other kinetic-energy weapons at the expense of research on exotic particle-beam weapons is likely to intensify as the administration looks for early results.

The fact that more tests will be necessary is sure to bring the administration back into conflict with Congress over the interpretation of the Anti-Ballistic Missile Treaty.

In the non-SDI research programme, the 'Orient Express' National Aerospace Plane is budgeted for a large increase in a joint project with NASA. Another collaboration with NASA is the Advanced Launch System, an expendable rocket for heavy payloads.

A set of basic research programmes in electronics is also emphasized, including a very-high-speed integrated circuit project and support for Sematech, the private microchip research consortium.

Other Agencies

The administration request for the US Geological Survey (USGS) reflects reductions for "numerous lower priority program activities". The areas being cut back include regional aquifer analysis, nuclear waste and coal hydrology and work on coastal erosion. Extra money is requested for advanced cartographic systems.

The National Oceanographic and Atmospheric Administration continues its fall from financial favour in the White House budget. Marooned by bureaucratic structure in the Department of Commerce, it will be cut in all research and operational areas, except environmental satellite data, which receive an increase mostly for the purchase of replacement weather satellites. □

Army renews controversial plans for biological warfare laboratory

Boston

THE debate over the US Army's plans to build a controversial high-containment facility to study biological warfare agents was rekindled with the recent release of a draft environmental impact statement attesting to the safety of the venture.

The Army's lengthy report concludes that the facility, designed for aerosol testing and experimentation with deadly and highly infectious biological warfare agents, "would have no reasonably fore-

design of the facility and in a programme to immunize workers against the diseases they will study. A two-month comment period now follows the issuing of the draft report, after which a final impact statement will be released. But critics of the Dugway facility are already complaining.

The proposed "maximum-containment" facility at Dugway has been a cause of controversy since its conception. The Army's current environmental impact statement emerged only after a court ruling in 1985 in response to a successful lawsuit brought by the Foundation on Economic Trends, led by Jeremy Rifkin.

Rifkin calls the latest report "grossly inadequate", claiming that the reviewers failed to consider seriously the possibilities of sabotage and terrorism, the proposed use of non-lethal simulants instead of actual pathogens, or the possibility of siting the facility elsewhere. Rifkin threatens to sue again unless these and other criticisms are dealt with in the final report.

Barbara Rosenberg, a research scientist at the Sloan Kettering Institute in New York and also an outspoken critic of the facility, and other scientists, including Nobel laureate David Baltimore, have argued against the construction of the facility at Dugway.

Very few known warfare agents require the highest biosafety level (BL-4) containment that the Army proposes for the Dugway facility. In fact, the Army claims that only research requiring the lower BL-3 rating will be conducted at Dugway, but it wants the capability to conduct the more dangerous research.

Seth Shulman



seeable impacts on the unprotected populace or the environment". The multi-million-dollar facility is at the Army's Dugway Proving Ground in Utah.

The Army's impact statement does acknowledge a potential hazard from "accidental occupational exposure" if an infected worker were to leave the facility, but claims that sufficient steps have been taken to avoid such a prospect in the

NIH conclude misconduct investigation

Washington

AN investigation by the National Institutes of Health (NIH) has concluded that there were "serious misrepresentations" in two published papers on the T-cell-activating substance interleukin-4A (IL-4A), and has recommended that the senior author of the paper, Claudio Milanese, be excluded from receiving any further money from NIH.

Milanese was a postdoctoral fellow in the laboratory of Ellis Reinherz at the Dana-Farber Cancer Institute in Boston. In October 1986, Reinherz became aware that some of the data provided by Milanese had been fabricated (see *Nature* 324, 193; 1986). He contacted NIH about the irregularities, and retracted two papers containing the fraudulent data (*Science* 231, 1118; 1986 by Milanese, Neil E. Richardson and Reinherz and *J. exp. Med.* 163, 1583; 1986 by Milanese, Robert E. Siciliano, Reinhold

Schmidt, Jerome Ritz, Richardson and Reinherz). Two other papers in the press were withdrawn. An investigation by Dana-Farber and Harvard Medical School concluded that, in addition to inventing data, Milanese had supplied fraudulent materials to others, and was guilty of "tidying" data to achieve better-looking results.

NIH have now accepted the conclusions of the Dana-Farber investigation. Although the main blame is laid on Milanese, the investigation noted that Reinherz's extensive research commitments "raised troubling questions about the supervision of junior staff in the laboratory." Harvard Medical School and Dana-Farber have agreed to form a committee to monitor Reinherz's laboratories. NIH praise Reinherz and Dana-Farber for investigating this incident and promptly informing the scientific community.

Joseph Palca

Do research centres herald a new era in UK academic science?

London

THE attention of Britain's scientific community is sharply focused on the new interdisciplinary research centres (IRCs) that will soon be sprouting up in British universities. Details of how the centres will be managed began to emerge last week. At the end of last year, the Science and Engineering Research Council (SERC) announced that the first centre, in high-temperature superconductivity, would be established at Cambridge. Earlier this month, three further centres were announced, in surface science (to be based at the University of Liverpool), molecular sciences (Oxford) and engineering design (Glasgow). A fifth centre, in synthesis and characterization of semiconductor and novel materials, will be set up at Imperial College, London, when more money is available. The Medical Research Council has expressed interest in jointly supporting the Oxford centre, as well as establishing a separate centre in toxicology. The Agricultural and Food Research Council and SERC have been discussing a joint centre in process engineering.

The importance of the IRC in shaping the future direction of academic science in Britain will depend largely on the performance of the first centres. Critics of the concept claim that directed programmes of research in large centres, often dictated by the needs of industry, will stifle initiative and are bad for science. The research councils, however, see the centres as a way of breaking down the barriers of traditional discipline-based research and encouraging vital collaboration and awareness between disciplines.

The proportion of the science budget devoted to IRC-based research is likely to increase markedly over the coming years. The Advisory Board for the Research Councils, which is largely responsible for the allocation of the science budget, said in recent advice to the government that "a large part of the councils' support for university research should eventually be channelled through these centres". The government's chief scientific adviser, Mr John Fairclough, has on several occasions made clear his support for IRCs.

Of the four centres announced, the surface science centre at Liverpool will receive the lion's share — £10 million over the next six years. This compares with £1.3 million that surface science at Liverpool received between 1982 and 1987. The Oxford centre expects to receive between £6 and £8 million, and the Glasgow centre around £7 million.

The centres will be managed at two levels. A local management committee

will administer the daily running of the centre. Membership of the management committee will in most cases include representation from the centre, SERC and industrialists involved in any of the research programmes. Overseeing the work of the centre will be a steering group, chaired by a member of SERC and with independent industrial and overseas representation. The centres will be reviewed

after two and four years. The fate of a centre will be largely determined by the steering group.

The extent of industrial involvement in the centres is likely to be considerable — although SERC has suggested that in the interests of a balanced research programme, income from industrial sources should remain at around 20 per cent of the total. Supporters of the concept are convinced that IRCs will provide the route to revitalize British research, but that the success of the venture hinges on adequate financing and flexible local management to foster originality. **Simon Hadlington**

New reform of Soviet education endorsed by Mikhail Gorbachev

London

MAJOR changes in the pattern of Soviet education advocated by the Central Committee of the Communist Party of the Soviet Union have been endorsed by Mr Mikhail Gorbachev. This is the third reform of Soviet education to be offered since Gorbachev came to power exactly three years ago.

According to Gorbachev last week, the new objective is a "qualitatively new level" of education in the Soviet Union. Mr Egor Ligachev, a fellow-member of the Politburo, says that the earlier reform of vocational education, introduced in 1985, had been prepared under the previous administration and was inadequate for the needs of *perestroika*.

One difficulty has apparently been that the reform had provided older secondary-school students with "mass vocational training . . . in the simplest specialities", reducing the entry of students into the last two years of secondary school and, eventually, into tertiary colleges. This, said Ligachev, runs counter to the need for ever-higher qualifications in the workforce.

There is some concern at the standards attained by school-leavers. A spot-check carried out last September among 27,000 first-year students at tertiary establishments of different kinds is said to have found 25 per cent of university students, and 45 per cent of those at technical colleges, to be below minimum standards; at one engineering college, 70 per cent of students failed the mathematics test administered.

Under the proposed reform, schools and courses for talented students will be encouraged, as well as a system of specialized boarding schools attached to universities and sharing their facilities on the model of the physics and mathematics boarding schools attached to the University of Moscow. Among other elements of the proposed reform are:

- Schools will be allowed to drop the

compulsory work-training previously required on the grounds that they lack the necessary time, facilities and people.

- Industry will be more closely involved in the financing of higher education and the tailoring of graduate output to job opportunities. The practice by which industrial organizations are paid 3,000 rubles for each graduate they recruit should, Ligachev noted, allow universities to collect 300 million rubles a year for the modernization of obsolete buildings and equipment.

- Universities and polytechnics will be encouraged to follow the examples of the University of Gor'kii and Tomsk Polytechnic in setting up industrial units, while the government should arrange for closer collaboration between higher education and the academies of science and the production ministries.

- Steps will be taken to free education from the "labyrinth" of rules and regulations, while institutions would be provided with better equipment and personnel.

One of the present difficulties seems to be that education rates only a low priority from the State Planning Commission, which tucks it away under "other consumers" in its budgets. Ligachev also complains about the design and condition of the buildings programme, citing the case of the Perm Polytechnic Institute, where an extension has been under construction for 20 years.

Another complaint is that too many tertiary institutions have poorly qualified staff. A particular difficulty seems to be that local authorities apply to the centre for permission to have a university in their region but then, when the institution is opened, wash their hands of it. Ligachev is for a "rigorous" survey of the Soviet Union's tertiary institutions, with a triage solution: borderline cases should be given help, but the weakest institutions should be made into local branches of stronger institutions. **Vera Rich**

Australia chooses spaceport consortium at Queensland site

Sydney

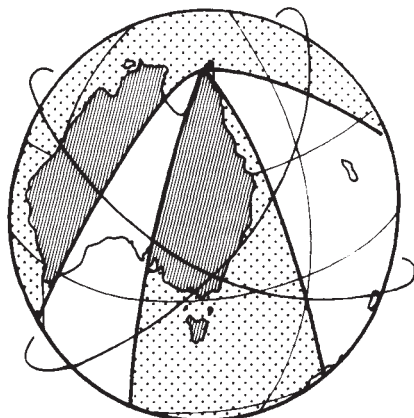
PLANS for a private enterprise spaceport on Australia's northernmost tip are winning backers. Last week, the Queensland state government announced that the Cape York Space Agency (CYSA) has been chosen to coordinate efforts to build a full-scale equatorial launch site by 1993.

A two-year, A\$5 million feasibility study will be the agency's first task. Aubrey Behn, chairman of CYSA, hopes it will end in agreements with both government and private organizations that will permit construction to begin.

Cape York has several advantages: it is near the Equator, where commercial launches are most economical, it is in a country with a stable government, and it has good communications, good weather and, above all, plenty of land. With three

years for construction, and an investment of A\$1,500 million, the first full-scale launch could be in 1993. A further A\$5,000 million would be needed to build additional launch-pads and support facilities to complete the spaceport in the year 2000, by which time Behn sees it supporting a township of 40,000 people.

CYSA is a consortium of 64 companies, most of them Australian, but including the Japanese construction company Shimizu Corporation. Companies with significant experience in rocket launching are notably absent from the consortium, although several had joined other groups that had hoped to be chosen to coordinate



The south-west launch sector from Cape York gives access to polar launches and includes sparsely populated areas around Alice Springs and the old launch site Woomera. To the east and south-east, the sector is over the South Pacific, with New Caledonia 2,500 km down-range.

Change for British Civil Service labs

London

RADICAL changes in the organization of the British Civil Service affecting much British civil science are likely to follow from a report of the government's central efficiency unit, published last week. The chief recommendation, which the government has accepted, is that units of the Civil Service providing public services should be regarded as candidates for the status of independent agencies, which in due course would be managed as units separate from the Civil Service proper.

Among the prospects for early formation into agencies are the Meteorological Office and the six non-nuclear research establishments of the Ministry of Defence. But ministers emphasize that the list of putative agencies made public last week is only a first list, and that others will follow.

The objective of the change is more efficient management, to which end the heads of agencies will be allowed a degree of flexibility (still undefined) about questions such as the salaries of those who work for them. But it is intended that agencies will remain responsible to ministers, and subject to financial scrutiny by public and parliamentary auditors, for which reason it is not yet clear how far the proposed changes will affect people's working lives.

While there is no suggestion that units with the status of agencies will eventually be made autonomous as well, their definition will enable the departments from which they spring to regulate more tightly than now the budgetary frameworks in which they operate, which may make for greater comfort or the opposite. □

the project. A proposal by the Australian arm of Sumitomo Corporation, the giant Japanese trading company, to go ahead and build a spaceport was rejected as management and control of the facility would have passed out of Australian hands.

The Queensland government has hedged its bets by giving CYSA only coordinator status and requiring it to bring other companies into the project. The success of the spaceport depends on how well CYSA meets these demands.

The Soviet Union is said to be particularly interested in launching a rocket from Cape York as Australia could make its entry into the commercial satellite launching market more palatable to Western nations. The near-equatorial site would also double the boost their rockets receive from the Earth's spin, permitting a large increase in payload.

Charles Morgan

Correction

THE news item "Move to ban sex determination" (*Nature* 331, 467; 1988) was written by Radhakrishna Rao and Elizabeth Ebbert. □

AIDS envelope protein patent

Washington

A US PATENT was issued last week to Harvard University for the outer envelope protein of the HIV (human immunodeficiency virus), the virus causing AIDS (acquired immune deficiency syndrome). Myron Essex and Tun-Hou Lee of Harvard are seeking the patent based on their discovery that the gp120 envelope protein reacts strongly with antibodies from blood of people infected with HIV.

The patent contains 10 claims covering gp120 and all proteins that cross-react with it, so most companies involved in the production of AIDS diagnostics, vaccines and possibly even therapeutics will need a licence. Harvard has given exclusive licensing rights to Cambridge BioScience of Worcester, Massachusetts, and representatives of the company say they will sublicense the patent for a fee. Licences will be required only for companies earning profits on products covered by the patent: researchers studying gp120 will not be required to pay royalties.

Substances that occur naturally are not patentable in their native form, but purified forms of natural substances often are patentable. Such "composition of matter" patents are desirable in biotechnology because patents for products are easier to defend than patents for processes.

In addition to the purified natural form of gp120, the Harvard patent will also cover synthetic and recombinant gp120. The patent could even be interpreted to cover products containing the gp160 protein — the gp120 outer envelope protein plus the gp41 transmembrane portion — if the patent is defined broadly. Antibodies to gp120 and some of the most popular cell lines for producing gp120 are included in the patent.

It is questionable whether the two most widely marketed diagnostic kits for HIV infection — produced by Abbott Laboratories and DuPont — will have to pay royalties under the Harvard patent. Both kits use killed whole-virus proteins linked to an enzyme to detect antibodies raised against any HIV antigen. But many companies are developing kits to detect only antibodies to gp120, because antibodies are thought to form against the gp120 protein very early in infection.

Manufacturers of vaccines against AIDS will be the most likely licensees of the Harvard patent. Genentech and collaborations between Chiron and Ciba-Geigy, and Repligen and Merck & Company are developing vaccines based on whole gp120 or gp120 fragments. Cambridge BioScience is developing a gp120 vaccine with Institut Merieux of France.

Carol Ezzell

New rules for carcinogens

Berkeley

CALIFORNIAN voters' attempt to take regulation of toxic chemicals into their own hands is off to a troubled start.

On 27 February, Proposition 65 comes into effect, requiring blanket regulation of all suspected carcinogens and teratogens. But critics argue that many of the proposition's regulatory teeth have been drawn since it became law in November 1986.

The impetus for Proposition 65 grew out of mistrust of the federal regulatory process, which has identified more than 200 compounds as potential carcinogens, but fails to regulate most of them (see *Nature* 330, 200; 1987).

Proposition 65 provides for regulation of all listed chemicals occurring at levels above the "level of no significant risk". That is defined as concentrations predicted to cause one additional cancer in 100,000 individuals over a lifetime of exposure for carcinogens, and as one-thousandth of the level producing no observable effects for teratogens. The accused violator has to prove no significant risk, and the law includes a bounty clause that encourages people to expose violations by giving them a share in any fines imposed.

Since the proposition was passed, controversy has centred on the list of chemicals to be regulated. Environmental groups were angered that the original list contained only the 29 chemicals known to cause cancer in humans. Over the past year, a scientific advisory panel has added most of the more than 230 chemicals named as possible carcinogens or teratogens by the National Toxicology Program and the International Agency for Research on Cancer, based on animal studies.

Despite that victory, environmental groups failed to secure a requirement that warning labels be carried on all products containing potential toxins. Instead, state guidelines allow posters advertising a toll-free telephone number which consumers can call for product information.

Many products have been exempted, in defiance of environmental groups' protests. The state exempted naturally occurring carcinogens after food producers complained that 15,000 produce items would require warnings because of trace levels of naturally occurring arsenic and other compounds. And the state guidelines rule that any food, drug or cosmetic product in compliance with federal regulations automatically meets the law's no-significant-risk standard, and need not carry a warning. That, critics say, defeats the whole purpose of stronger local regulations.

Marcia Barinaga

Appeal for computer facilities in West German universities

Munich

THIS is a critical year for the research budget in West Germany. Funds to the West German Education Ministry for medium-term university support are due to be cut by 20 per cent, some of the *Länder* (states) have been hit hard by losses in tax revenue from ageing industries such as coal-mining and steel and final preparations are being made for a sweeping 1990 tax reform.

One result is that research organizations such as the Deutsche Forschungsgemeinschaft (DFG) and the Wissenschaftsrat have been raising their voices ever higher for the maintenance of public funding. The DFG is the primary supporter of basic research in West German universities, and the Wissenschaftsrat a council of researchers from all fields that advises the federal and *Länder* governments, always partners in research support.

In a January report sponsored by the DFG, a commission of 12 West German academics called for increased support for computing facilities in universities. Without it, says the report, "adequate research and teaching will be impossible". The commission is particularly afraid that West Germany will be left behind by the rapid upgrading of computer facilities at universities in other countries.

The report calls for investments of at least DM 400 million a year until 1991 for everything from mainframes to personal computers, as well as computer networks.

The report says that the largest single amount should go into scientific workstations (DM 83 million a year), with DM 60 million for mainframe computers and DM 60–65 million for the establishment and maintenance of computer networks.

The figure is significantly higher than what was called for in the 1988 "framework plan" for university investment, the latest of the Wissenschaftsrat's annual documents which is meant as a guideline for both federal and *Länder* governments on higher education support.

The document says that, to make good the disparity between West Germany and other countries, the expected cut in the West German Education Ministry budget must be restored and that even the north German *Länder* in the direst financial straits must submit more applications for computer facilities.

The triennial commission report, the third of its kind, may be well timed. All the science organizations can do, said the DFG's Alfred Küllmer, "is to make the politicians aware of the consequences" of budget cuts.

So far, West Germany has kept up with advances in computer technology elsewhere, says the report. But now there must be a push, as in the United States, to bring more students in more subject areas into contact with computers and to increase the use of computers in society at large.

Steven Dickman

Ministers act to reduce industrial accidents

Paris

ACCORDING to the Organisation for Economic Cooperation and Development (OECD), there are about 20 major accidents* involving hazardous (non-radioactive) substances each year in OECD countries and as many as 200 accidents "causing significant damage to health, environment or property". Over the past 14 years, such accidents have caused more than 4,100 deaths and almost 56,000 injuries. Against this background, environment ministers and high-level representatives of 24 OECD countries met in Paris on 9–10 February to propose new, legally binding guidelines on accident prevention and emergency measures.

To attack the anomaly that it is often a neighbouring country that bears the brunt of pollution, two council acts were put forward at the conference. First, mutual consultation arrangements were advocated for siting new installations in frontier regions. There should also be an exchange of information about the safety of existing

installations and the obligation to inform neighbouring countries immediately in the event of an accident. Second, the public was felt to have a right to information, both what to do in case of an accident and in relation to the prevention of risks.

If the OECD council approves the proposed acts at its meeting in the summer, they will immediately become legally binding. Not legally binding, however, is the recommendation of a "polluter pays" principle. In the event of a major accident, the company responsible may be sued for compensation.

There was unanimous agreement at the meeting that safety precautions should be the same irrespective of the economic status of the country. It was also recommended that OECD should link up with the United Nations Environment Programme and other international organizations to study this issue in greater depth.

Peter Coles

* A major accident is defined as causing at least 5 deaths, 25 people injured and/or evacuation of 400 people.

South African university politics still hanging in the balance

London

A FULL bench of the Cape Supreme Court has ruled that the conditions imposed on the universities of Cape Town and the Western Cape (UCT and UWC) are invalid and has ordered the ministers of Education and Culture in the Houses of Assembly and Representatives to pay the costs of the application. The conditions, promulgated on 16 October last year by the South African government, provide for cutting the state subsidy to universities refusing to report incidents of political activity on their campuses (see *Nature* 329, 192; 330, 4; and 331, 104, 1987).

The bench, comprising Mr Justice Nel, Mr Justice Conradie and Mr Justice Howie, made this ruling on 11 February; full judgement will be furnished at a later date. In a parallel application, counsel for UCT and UWC asked the ministers to show why the decision to impose the conditions should not be reviewed or set aside. Alternatively, they applied for an order declaring the conditions invalid and without force and effect. Evidence produced in court by the universities con-

tended that any substantial reduction in their annual subsidy would gravely prejudice the universities' educational functions, and render them unable to fulfill their purpose and obligations under their founding statutes and other relevant legislation. They submitted that the conditions encroached upon and were in conflict with the autonomy and functions conferred on the university by parliament.

The outcome of a similar application by Natal University is currently being awaited, but the Cape Supreme Court's ruling is not binding in the province of Natal. Depending on the outcome of the Natal application, the government may decide to take the case to the Appellate Division of the Supreme Court in Bloemfontein, but in the interim they are powerless to impose a subsidiary cut on either UCT or UWC without changing the law itself. While their huge parliamentary majority would allow them to do this, the Cape Supreme Court ruling provides the ideal opportunity for the government to back down, if it is inclined to do so.

Michael Cherry

Royalties to be paid on animal patents

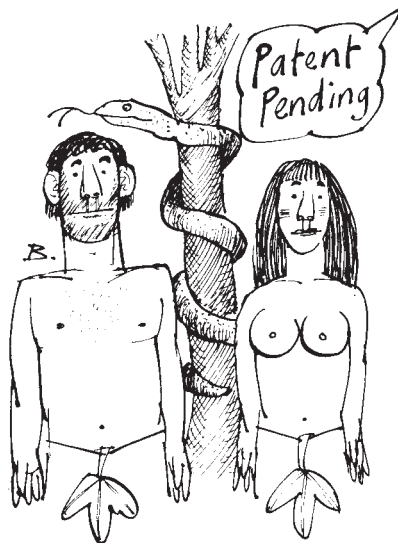
Washington

FARMERS who breed genetically altered animals covered under US patents will have to pay royalty fees, according to a letter from the US Patent and Trademark Office (PTO) released two weeks ago. The letter, written by patent office commissioner Donald J. Quigg to the chairman of the Senate subcommittee on patents, copyrights and trademarks last autumn, was released to the media by the Foundation on Economic Trends, headed by Jeremy Rifkin. In the letter, Quigg makes it clear that according to the current interpretation of patent statutes, "unauthorized acts of reproduction . . . [of patented animals] would seem to be an infringement".

The effect of animal patenting on farmers has been debated hotly since the PTO announced last spring of that it would consider applications for such patents (see *Nature* 326, 729; 1987). Major farming organizations have protested against PTO's action, and legislation was introduced by Senator Mark Hatfield (Republican, Oregon) in the last Congress to impose a two-year moratorium on animal patenting so that the economic and ethical implications could be resolved. Although the legislation was not passed by Congress, Hatfield plans to introduce it again this year. The congressional Officers of Technology Assessment is also preparing a study of the potential effects of the patent-

ing of animals, due for release this summer.

The PTO is currently reviewing 20–22 applications for patents on animals, and provided that the applications meet the



usual criteria for patentability, the first animal patent could be granted as early as this summer. But the ability to patent animals will not ensure market success. If genetically altered animals are not competitive in price or better than those produced by traditional methods, farmers may still choose to buy animals that will not be subject to royalty payments. Carol Ezzell

Rehabilitator for Vavilov comes to light

London

A MEMBER of the Soviet Academy of Sciences, Dr Artem Takhtadzhyan, last week publicly acknowledged his debt to the emigré writer Mark Popovsy, thus signalling a further relaxation of the rules of public discussion in the Soviet Union.

This development has come about in *Literaturnaya Gazeta*, in which there has been a vigorous correspondence since the centenary of the birth of the geneticist Nikolai Vavilov on 25 November last year. Popovsy, who emigrated from the Soviet Union in 1977, is an authority on Vavilov's life and the circumstances of his death (in prison) as one of the victims of the Lysenko period, but apart from one mention in a scientific bibliography last year, this is the first time Popovsy's work on Vavilov has been acknowledged.

Vavilov was formally rehabilitated in 1955, but open discussion of the purge of geneticists was permitted only for a brief period at the end of the 1960s. Popovsy then published an account of the campaign against Vavilov in the 1930s, evoking the atmosphere of fear preceding Vavilov's arrest in 1940, but only in the journal *Prostor*, published in Alma-Ata.

After leaving the Soviet Union, Popovsy pursued his work on Vavilov, gaining access to much archival material, including the secret police file on Vavilov housed in the State Prosecutor's Office; his findings were published in *The Vavilov Affair* (Archon Books) only in 1984.

In the Soviet Union, preparations for the Vavilov centenary began in 1985, when two films were commissioned. In November last year, *Ogonyok* published an article by the script-writer of one of the two films that stressed the scale of the persecution of the 1930s, while *Literaturnaya Gazeta* published a long article by Academician Takhtadzhyan which drew heavily on Popovsy's material, noting Vavilov's ambivalence towards Lysenko and that Vavilov, although at first denying the charges of sabotage and espionage against him, eventually "recanted", not necessarily because he was tortured.

Since last November, Soviet readers have written to ask how reliable these revelations are, but those elsewhere have noted with regret the use of Popovsy's material without acknowledgement. Takhtadzhyan has now answered both sets of critics by quoting the cooperation of the official Vavilov commission and saying that "the basic material on Vavilov's work and his tragic end were obtained due to the energetic activities of the writer M.A. Popovsy..." Vera Rich

Advising the government

SIR—John Maddox is correct (*Nature* 330, 515; 1987) in emphasizing the interdisciplinary complexity of arms control issues that has led to a need for “more intellectual adventurousness”; the very complexity of these issues indicates the need to create interdisciplinary centres such as Stanford University’s Center for International Security and Arms Control. In fact, the Stanford Center has grown with the conviction that a balanced mix of scholars, practitioners and students from many disciplines is needed to generate the new ideas and to train students in national security planning, of which arms control is one important part.

I regret, therefore, that Maddox suggests that I or my colleagues view the Stanford Center as one that simply “exists to replace the generation of natural scientists who have given valiant public service in the cause of arms control during the past three decades”. He badly misses the mark by focusing on only one aspect of our activities. National security planning is a field of political action and, as Albert Einstein once said, “politics is much harder than physics”. But when political leaders set policy goals and act, they must do so on the basis of informed technical data and projections of future technical possibilities. Policy goals cannot coerce laws of nature. President Reagan’s poorly conceived Strategic Defense Initiative (SDI) is a case in point.

It has become painfully clear in the subsequent debate that there is a need for a cadre of technical experts trained and utilized to provide an informed and independent input into such issues which, although essentially political, must rest firmly on an understanding of technological reality.

During the 1960s and into the early 1970s, the President’s Science Advisory Committee, through its working panels, helped to train such a cadre. Today, after a 15-year lapse, the United States needs a new generation of informed, technically expert and *independent* scientists to serve that function. The Stanford Center will be making a valuable contribution if it can help fill that need as one of its research and teaching activities.

More broadly, the United States needs to recreate an effective presidential science advisory mechanism. Many of the early unrealistic claims for SDI (some still occasionally heard) could have been avoided had such a mechanism been in place. Here again, Maddox misses the point in characterizing the debate of the past four years as lacking the “excitement of earlier decades”.

As one who has participated in both the 1969–72 and the 1983–87 debates on anti-ballistic missile systems, I disagree com-

pletely. There have been a number of independent, thorough and well-informed studies on strategic defence, involving a large number of academic and industrial scientists. These have been mostly carried on outside government and have played an important role in the public debate — witness the recent study by the American Physical Society, which clearly discusses the technical challenge of directed energy weapons for strategic defence.

Just because of the vacuum created by the demise of effective scientific advisory channels in government, independent studies have played a very important role in informing the SDI debate — and at a high level of scientific sophistication and intensity.

The recent Intermediate Nuclear Force agreement has created still more difficult technical challenges. Indeed, Maddox is right in saying that on-site cooperative measures of inspection are very difficult political issues. However, the need for scientists to contribute here will in no way be diminished. Unless Maddox knows something that I do not, it will take clever analysis if the verifiers are to be able to say with confidence and substance that specific agreements can indeed be verified.

SIDNEY D. DRELL

*Stanford Linear Accelerator Center,
Stanford University, PO Box 4349,
Stanford, California 94305, USA*

Directing science

SIR—In the many letters written about the Advisory Board for the Research Councils (ABRC) report of July 1987 on *A Strategy for the Science Base*, there has been little discussion about which people or organizations should be involved in the future direction of science. Most of the discussion has been focused on *what* those directions should be and *where* the research should be concentrated. In its written response to the ABRC, the Institute of Mathematics and its Applications (IMA), among other comments (to be published in the *IMA Bulletin*), made a point of first welcoming the open discussion by ABRC, and then asking it to pursue this policy of openness in future. Policy decisions from the strategic down to the lowest grant-giving levels in the research councils are arrived at largely in secret and with very little open discussion before or after the decisions. For example, new policy instructions by the Science and Engineering Research Council mathematics committee are never discussed with mathematical societies, and the same goes generally (but not always) for other committees.

The IMA strongly believes that the scientific, technological and engineering

institutions should be involved in consultations both before and after changes in science policy. Such consultation need not be lengthy, but should be an open process.

Most outside advice is at present given by individuals: there is little encouragement for organizations to discuss internally these policy questions and then give advice *as* organizations. Concerned advice given by a group may (though not always) contain more and better ideas than those of individuals. As UK scientific and technological organizations, such as the IMA, have a wide membership in research, industry and government establishments, their voices should be heard. They could provide a useful and innovative input as, for example, the institute’s response to the Advisory Council for Applied Research and Development where advice on computer verification was given. They could make a particularly valuable contribution in multidisciplinary policy questions.

The IMA believes that it and other scientific institutions could play a useful role in informing their own members in academic life, industry and government establishments about science policy decisions and their implementation. The institutions already play a valuable role in organizing conferences, workshops and short courses on subject areas that the research councils are encouraging, and this could be increased.

R. F. CHURCHHOUSE
(President)

*The Institute of Mathematics
and its Applications,
Maitland House, Warrior Square,
Southend-on-Sea, Essex SS1 2JY, UK*

Time for secrecy

SIR—In his review of Geoffrey Cannon’s book *The Politics of Food*, John Rivers (*Nature* 238, 769; 1987) suggests that in Britain the time of day will soon be an official secret. I am not sure that this hasn’t happened already. Precise time-keeping has become one of the cornerstones of modern military technology. In the United States for instance, the Naval Observatory has responsibility for time-keeping and serves civilian government agencies as well as its military. But very few civilian agencies ever need to know the time of day with extreme precision and it is not clear whether the military has reserved the most precise measurements for itself as has happened in the case of the military Navstar global positioning system. Even if the time of day is not yet an official secret in Britain, where we all are is information which the Pentagon has classified.

PETER WILLIS

*Department of Physics,
University of Auckland,
New Zealand*

The death of Richard Feynman

Richard Feynman was rightly an heroic figure in physics for the best of reasons: he had the flair to make up the rules as he went along. Will young people have heard the message?

THE death last week of Richard Feynman of the California Institute of Technology is not merely a loss to science but a setback for the principle that science can (and should) be fun. That was the spirit in which his distinctive physics textbook is decorated on the flyleaf with a photograph of him at work on a the bongo drums, but there is much more than that to say.

Feynman's carefully preserved Brooklyn accent allowed him to make even his acerbity amusing, while his pleasure at his many achievements was usually modified for the benefit of observers who might be dazzled by his readiness to cut the figure of a poor boy made good. He was always well-supplied with plausible accounts of how he stumbled into being a genius.

There will be many formal appreciations of Feynman's work in the months ahead. (*Nature* hopes to publish such a one.) What follows is not that at all, but one view of how Feynman helped to change the way that people think from the early years after the Second World War (Feynman, with others, had been at Los Alamos working on the Manhattan Project) when it became everybody's task to make a sensible quantum theory of electrodynamics, the task from which Dirac had shied away.

There were two difficulties, of which the first was conceptual: whole generations had by then been brought up to understand classical fields such as those of Maxwell's electromagnetism, but what precisely was to be understood by the notion of a field which, like Maxwell's, was not merely a function of place but, to be relativistic, of time as well and, being observable, capable of being represented as an operator, as quantum mechanics requires?

The practitioners in those far-off days did not seriously appreciate what they required of those who listened to them. Perhaps they were forgivably distracted by their second difficulty, that the quantization of the electromagnetic field did not magic away the classical conundrum first made explicit by H.A. Lorentz, that the energy of the interaction between a charged object and its own field must be infinite. Dirac had had better luck with his theory of the singlet electron.

The Schwingers and Tomonagas of the times, sharing the general belief that the infinities must be artefacts, hit the problem with all the algebra they could lay their hands upon. It was a triumph of

application (not to mention high skill) over great adversity. Self-mockingly simple, Feynman once said he merely wanted properly to understand what the giants had been doing.

One of the reasons why Feynman was always a physicists' physicist is that he really meant that he wanted to understand. The most memorable of the by-products of that exercise (while still a graduate student, but an exceptional one, at Cornell) are the Feynman diagrams, all-purpose ways of visualizing, classifying and calculating the elements that go into calculating the probability that a system will be transformed into another upon interaction. (Matrix element is the keyword.) Undergraduates who cannot scribble a few Feynman diagrams will not now easily succeed. But that, the prevalence of Feynman diagrams, and the plague of their corruptions, would be a poor monument for a subversive man.

Feynman's enduring achievement was rather that he made the quantum mechanics of fields accessible to others than people skilled at algebra. Starting from a point of view of which Bohr would heartily have approved, he took the view that a quantum system is no different from a classical system in that one thing follows from another. There may be a little local difficulty, not insuperable, in shaking this notion of causality free from that of the expectation that time must be an absolute simulation of some unique clock-face.

What really matters about the evolution of a physical system, Feynman concluded at a tender age, is that it should be possible to calculate the evolving state of any system from a description of its present state in an appropriate language as modified by some operation, mathematical if the language is of that kind. Feynman's propagators (which are the ways of calculating) will ultimately be more important than his diagrams.

More important still is Feynman's way of getting to this state of grace. The textbooks are full of conundrums due to Bohr in the 1920s based on his version of the Young's slit *gedanken* experiment: put an electron source in front of a pair of slits, stop down the output of the source until the electrons are emitted, for practical purposes, singly and then ask why things should be different if one slit is closed.

Feynman appears to have been undismayed by mean *gedanken* experiments of that kind: "if it's a quantum system, all

evolutionary tracks are possible, but it's important to be able to estimate the likelihood of one rather than another". So there emerged Feynman's distinctive formulation of the way that we all now conceive of the deterministic evolution of systems that are no less well-determined for being intrinsically uncertain. That is like pulling two rabbits out of a hat at once. That the infinities melt away is another.

Most of this had been accomplished in the early 1950s. Feynman became perpetually Caltech's most popular lecturer and also (a different business) one of its best teachers, even at such a place. All that was accomplished in about fifteen years.

There are several lessons to be drawn from the Feynman story, of which the chief concerns the nature of the scientific life: there is no cant in the common assertion of great men that they are still learning, even from their students. The name of the game is to understand: to have understood a little better is the prize. But understanding does not require that one should follow the other fellow's algebra in every detail, but that one should look for a language that makes sense. That is what the true subversives practise.

Second, the time quickly comes when it is more economical of genius to teach rather than to practise. It is simply that there is a limit to the energy there can be to spare for the usual competition for space in established journals, and a natural pleasure and evident benefit (if to others) to be won from teaching people in a way that makes them properly subversive also.

Third, there is no end in sight. Feynman's account of quantum mechanics and of the way that quantum systems evolve is in a sense provisional. It may be the best system extant, with applications way outside the quantum electrodynamics for which it was developed, but that does not imply that some young subversive will not one day find an even simpler way of putting the whole message, as Feynman did. It will be a great surprise if one does not. But the need for another simplification is already clamant.

From time to time, candidate successors emerge, but only briefly. Feynman's strength was his physical intuition, his weakness was his flair for being distracted. Einstein, of course, was more persistent, but that was a less distracting age.

John Maddox

Fluid mechanics

Models of jovian vortices

Andrew P. Ingersoll

THE Great Red Spot of Jupiter is a huge, circulating mass of fluid. It is probably more than 300 years old¹, which is remarkable because the atmosphere in which it sits changes its appearance at smaller scales every few days^{2,3}. The large-scale flow organizes the cloud's chemical composition, giving the spot its reddish colour. Early theories^{1,4} held that the cloud feature was tied to a solid object floating in the ocean or resting on the solid surface. Later it became clear that Jupiter has no solid surface except perhaps for a small rocky core within 15 per cent of the planet's radius⁵. Recent theories⁶⁻¹⁰ have focused on fluid-dynamical questions. What properties of Jupiter's atmosphere allow such flow features to exist? In general, do large vortices form spontaneously in turbulent fluids? Can simple models of the Great Red Spot be constructed on the computer or in the laboratory? According to Marcus⁶ and Sommeria, Meyers and Swinney⁷ in their papers⁶ on pages 693 and 689 of this issue, respectively, the answer to the last question is yes.

The Great Red Spot extends 25,000 km in longitude, one-third the radius of Jupiter (Figs 1 and 2). Smaller spots circle it in 7 days, or 17 times the planet's period of rotation. The spot sits in an anticyclonic shear zone, rolling like a ball-bearing between a westward current to the north and an eastward current to the south. These properties — the relative size and period, the relationship to the nearby currents, and the long life — obviously must be matched in any successful model.

Neither numerical nor laboratory models can encompass the full range of scales in a planetary atmosphere. Atmospheric flows are turbulent and inviscid on scales from 1 to 10^7 m. In a laboratory fluid, viscosity limits small-scale motion, and resolution in a numerical model is limited by the computer. In both cases, the goal is to retain enough of the small scales so that all the important properties of the flow are represented. Another limitation is ignorance about the lower

boundary conditions and energy sources. Even the Voyager spacecraft did not directly measure the winds or temperatures beneath the clouds. Indirect methods¹¹, using the Voyager data in combination with dynamical constraints, provide only a partial look at the deeper levels. Most authors assume that the Great Red Spot is a shallow feature relative to the depth of Jupiter's atmosphere.

In his paper in this issue⁶, Marcus numerically integrates the equations of

breaks up. He also shows mergers, in which a large spot grows by sweeping up smaller spots with the same sign of potential vorticity. The impression is of large-scale order spontaneously emerging from small-scale chaos. Snapshots of the film are shown on page 694 of Marcus's paper.

Sommeria *et al.* in their paper in this issue⁷ study the flow in a rotating cylindrical container. They force a zonal shear flow by pumping fluid into a tank at one radius and out of the tank at another radius. The Coriolis force (an apparent force to the right of the wind in a system rotating anti-clockwise) on the outward radial flow produces a broad zonal current or jet. Viscosity is relatively unimportant for these large-scale motions. Above a certain pumping rate, the jet is unstable

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

Fig. 1 Left, the Great Red Spot of Jupiter as photographed by Voyager 1 from 35.6×10^6 km. (Courtesy of Science Photo Library and NASA). Right, a stable vortex created in the simulation by Sommeria *et al.* Red dye has been added to reveal the flow lines (Courtesy of H. Swinney).

motion for a single layer of fluid with various initial conditions and lower boundary conditions. Viscous dissipation is neglected so the flow conserves mass, angular momentum and energy. The most impressive output of his numerical experiment is an animated film in which fluid parcels are coloured according to their potential vorticity, which is proportional to their angular momentum about a vertical axis and is a constant for each parcel. (As fluid columns change depth or latitude, their local rate of spin, or vorticity, changes according to a simple formula. The constant in this formula is the potential vorticity.) The film shows examples of unstable behaviour, where a vortex that is initially rolling the wrong way with respect to the ambient shear flow

and breaks up into one or more vortices with many of the characteristics of the Great Red Spot. Single isolated vortices form in the strongest jets (Fig. 2). When more than one vortex is initially present and the current is strong enough, the vortices merge as they do in Marcus's numerical experiments. Merging is also common on Jupiter¹². Both the numerical model and the laboratory experiments leave little doubt that large vortices can exist in chaotic zonal shear flows.

An important point made in both papers^{6,7} is that mixing tends to produce a state of almost-uniform potential vorticity. Non-uniformities are gradually swept up into isolated patches like the Great Red Spot. The patches cannot radiate away their energy because Rossby

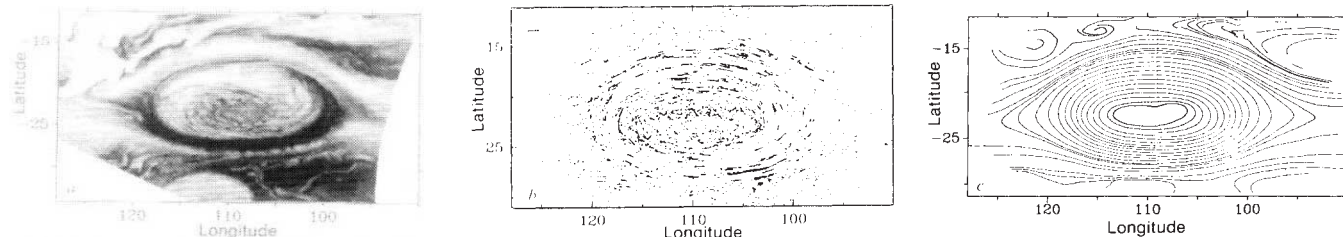


Fig. 2 *a*, Voyager 2 image of the Great Red Spot with planetographic latitude along the ordinate and longitude along the abscissa (courtesy of NASA); *b*, measured velocities with line length indicating wind speed (scale bar, 100 m s^{-1}), and line direction pointing downwind from the dot; *c*, fluid trajectories computed¹¹ from the measured velocities with dots indicating intervals of 10 h.

waves, the low-frequency oscillations of a rotating fluid, cannot propagate. (Just as waves in the ocean require a density in which to propagate, Rossby waves require a potential vorticity gradient.) This tendency could be the general mechanism by which isolated vortices maintain themselves. Jupiter's atmosphere is represented by Marcus's Fig. 3 on page 694, which has several alternating eastward and westward jets. For this case, uniform potential vorticity in the visible clouds implies that the underlying layers are in differential rotation — a conclusion reached independently by our recent analysis of Voyager data¹¹.

The next step is to gain some experience with these models. There are now several apparently successful models of jovian vortices⁶⁻¹⁰ which differ mainly in the assumed boundary conditions and energy

sources. Some of the assumptions may turn out to be incorrect when applied to Jupiter's atmosphere. All the models, however, are valuable because they help to define the range of conditions under which stable, isolated vortices can exist. □

1. Peek, B.M. *The Planet Jupiter* (Faber & Faber, London, 1958).
2. Smith, B. *et al. Science* **204**, 951–972 (1979).
3. Smith, B. *Science* **206**, 927–950 (1979).
4. Hide, R. *Nature* **190**, 895–896 (1961).
5. Stevenson, D.J. & Salpeter, E.E. in *Jupiter* (ed. Gehrels, T.) 85–112 (University of Arizona Press, Tucson, 1976).
6. Marcus, P.S. *Nature* **331**, 693–696 (1988).
7. Sommeria, J., Meyers, S.D. & Swinney, H.L. *Nature* **331**, 689–693 (1988).
8. Ingersoll, A.P. & Cuong, P.G. *J. Atmos. Sci.* **38**, 2067–2076 (1981).
9. Read, P.L. & Hide, R. *Nature* **308**, 45–48 (1984).
10. Williams, G.P. & Yamagata, T. *J. Atmos. Sci.* **41**, 453 (1984).
11. Dowling, T. & Ingersoll, A.P. *J. Atmos. Sci.* (in the press).
12. Mac Low, M.-M. & Ingersoll, A.P. *Icarus* **65**, 353 (1986).

Andrew P. Ingersoll is in the Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, California 91125, USA.

Epidemiology

HIV infection in heterosexuals

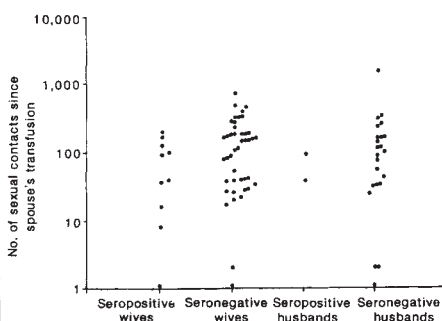
Robert M. May

MANY aspects of the epidemiology of AIDS (acquired immune deficiency syndrome) are not yet well understood, partly because infection with HIV (human immunodeficiency virus) is usually asymptomatic, with the disease appearing only after a long and variable incubation period. In particular, the asymptomatic nature of most infections makes it hard to estimate the probability that an infected individual will transmit HIV infection to a susceptible partner within any specific risk-group, resulting in uncertainties about how likely it is that a slower, 'second-wave' epidemic of HIV/AIDS in developed countries is spread by purely heterosexual chains of transmission¹. Help may be at hand in the form of a paper by Peterman *et al.*², who use data gathered from the heterosexual partners of adults infected with HIV by blood transfusions (so the date of the original infection can be assigned), to obtain estimates of the corresponding male-to-female and female-to-male transmission probabilities. Most interestingly, their study also suggests that the probability of transmission is not dependent on the number of sexual contacts that a given couple had.

From a larger pool of patients with transfusion-associated AIDS, Peterman *et al.* selected 88 infected individuals and their spouses for the study and added 18 couples where the index case showed serological evidence of HIV but had not developed AIDS. Of these 106 index cases, 26 had had no sexual contact with their spouse since transfusion, leaving a pool of 80 infected individuals — 55 men and 25 women — and their spouses. Peterman *et al.* conducted standard interviews with the

index cases and their spouses and other family members, reviewing the types of contact with the HIV-infected family member, medical and social histories, and other possible exposures to HIV (the study excluded individuals with other risk factors). Wives/husbands were asked about their sexual practices and about the average number of sexual contacts with their infected husband/wife per month since the transfusion; they were also asked about changes in frequency of sexual contacts after the partner had been diagnosed as having HIV seropositivity. All family members were tested for HIV antibodies.

Peterman *et al.* find that 2 of the 25 husbands of infected wives, and 10 of the 55 wives of infected husbands, were HIV seropositive. The figure shows the



For each of 80 individuals, this figure shows the number of sexual contacts with a spouse, after the spouse was (unknowingly) infected with HIV by transfusion. The 25 husbands and 55 wives of infected individuals are divided according to current serological status. In this study group, transmission probability is not correlated with number of sexual contacts. (After ref. 2.)

relationship between transmission of HIV infection and the number of sexual contacts with the infected spouse, and suggests the surprising conclusion that transmission probability is unrelated to the actual number of sexual contacts, for partnerships ranging from one to several thousand contacts. Insofar as there is any statistical correlation between transmission probability and number of contacts, it is negative (although not to a statistically significant degree): one infected woman had had only a single sexual contact and another had only 8; 11 of 55 wives and 5 of 25 husbands remained uninfected after more than 200 sexual contacts with their infected spouses.

Some details of the study of Peterman *et al.* deserve mention. First, their 80 index cases comprise 66 with AIDS and 14 with asymptomatic HIV infections; the authors justify pooling the two categories on the reasonable grounds that the transmission probabilities are not significantly different (10 of 66 spouses infected in one category and 2 of 14 in the other). Second, interviews with the 12 seropositive spouses of index cases suggested that only 1 of the 12 had sexual contact with anyone besides the spouse since 1978 (a woman who reported two other partners, neither in a defined risk-group). Third, this woman was also the only seropositive spouse to report any history of other sexually transmitted infections (gonorrhea); two seronegative spouses also reported a previous sexually transmitted infection (genital herpes simplex). Fourth, the 10 wives who acquired infections tended to be older (median age 62 years, range 27–73 years) than the 45 wives who did not (median age 54, range 25–72 years). This age distribution is a reminder that transfusion-associated HIV infections may be atypical; recipients of blood transfusions tend, on average, to be older and sicker than the population at large. All interpretations of the results² are subject to this reservation.

Some mathematical models for the transmission dynamics of the disease assume the transmission probability is, on average, constant for all partnerships of a given kind, independent of the duration of the partnership or the number of sexual contacts within it^{1,3,4}. This is, for example, the assumption we made¹ in defining our transmission probability β (although we may have caused some confusion by not making this definition as clear as we might have). Other models^{3,5} assume a constant transmission probability, γ , per sexual contact, so that the overall probability for an infected individual to infect a sexual partner increases with the number of contacts (giving on average a transmission probability per partnership of approximately $1 - \exp(-\gamma n)$ after n contacts). The figure suggests that, at least for the data collected by Peterman *et al.*, the first assumption — some average transmission

probability per partnership — is closer to reality than the second. These data are grossly inconsistent with any Poisson process, whereby some constant transmission probability per contact compounds independently randomly. But we must be cautious in extrapolating from this possibly atypical data set, and it seems reasonable to assume that, in general, transmission probability may depend to some degree on the duration of a partnership and/or the number of contacts.

Peterman *et al.* consider various explanations for the lack of dependence of transmission probability on number of sexual contacts in the figure. As summarized in the table, they find no significant differences in the types of sexual contact of the seropositive and seronegative spouses. There is some indication that infectiousness varies with time, being higher in the first few months after infection is first acquired and again at the onset of AIDS as the immune system collapses,

estimates. As reviewed by Peterman *et al.* these include studies of HIV infection among wives of infected haemophiliacs (10/148 in one study¹¹, and 2/21 in each of three other studies¹²⁻¹⁴), and of female sexual partners of bisexual males (8/43)¹⁵. Higher transmission probabilities have been reported for female sexual partners of intravenous drug abusers who claim not to be drug abusers themselves (32/76)¹⁶; for spouses of infected US military personnel (5/7)¹⁷; and for wives of AIDS patients in Miami (14/28)¹⁸ and in Zaire (11/18)¹⁹. In all these studies, however, sexual partnership with a single infected individual (with neither partner having any other sexually transmitted infection) seems to be not the only risk factor. Moreover, none of these previous studies gives information about the relation between transmission probability and number of sexual contacts in the way the figure does.

What does seem clear is that HIV has a lower transmission probability than other

number of family members tested and found to be seronegative now totals 442 Many additional family members in this country have now been tested by their family physicians, yet no one has been reported to the Centers for Disease Control who appears to have acquired an infection by a casual family-type contact." On the other hand, there is one report²⁵ from Germany of horizontal transmission of HIV to a sibling of an infected young child. By the same token, although most accidental cases of HIV infections among hospital workers derive from needles, there are cases which appear to arise from blood or serum inadvertently coming in contact with chapped hands or faces. The chance of HIV transmission by casual contact is clearly very, very low, but it is not strictly zero (although it may be zero for all practical purposes).

When it comes to placing an individual infected with HIV in a school or workplace, we should, in principle, multiply the infinitesimal probability of transmitting infection by the number of potential infectees, each with a normal life expectancy, and weigh this against the shortened life expectancy of the infected individual. This would be an anguished calculation, even if we knew the probability of transmitting infection by casual contact (10^{-4} ?; 10^{-7} ?), which we do not. It is paradigmatic of many similar social and political calculations that must deal with events that have a very low (and uncertain) probability of occurring, yet whose occurrence would be awful; what might be called 'zero times infinity' problems. Difficult though such problems are, I do not think it helps to define them out of existence by asserting that some event of very low probability simply cannot occur. □

Types of sexual contact by spouses of patients with transfusion-associated HIV infection

	<i>n</i>	Vaginal intercourse	Anal intercourse	Fellatio	Cunnilingus	Ever used condoms
Husbands						
Seropositive	2	2	0	0	0	0
Seronegative	23	23	1	6	4	2
Wives						
Seropositive	10	10	1	2	2	0
Seronegative	45	45	1	11	12	4

than at intermediate times in the incubation period⁶⁻⁹. This would result in sexual contacts at some times being much more likely to transmit infection than those at others, but I find it hard to see how this could explain the patterns in the figure. As mentioned above, essentially none of the spouses had a history of other sexually transmitted infections, so that 'cofactors' of this kind can be excluded here. Peterman *et al.* observe that it could be that some infected individuals are more infectious than others (for instance, shedding more virus) and/or that some spouses are intrinsically more likely to acquire infection (for physiological or genetic reasons). This could explain the patterns but, understandably enough, there is no evidence bearing on either of these possibilities. One exception is the suggestion that a group-specific component allele *IF* may be associated with enhanced susceptibility to HIV infection¹⁰. Peterman *et al.* tested this suggestion, but find that the allele was not present at significantly different frequencies for 10 seropositive spouses than for 60 seronegative spouses (specifically, the frequency of the *IF* allele is 0.15 among the 10 infected spouses and 0.20 among the 60 uninfected ones).

Given the small sample sizes, the estimate of Peterman *et al.* of an average transmission probability of around 0.1 (2/25) female-to-male and around 0.2 (10/55) male-to-female is consistent with other

heterosexually transmitted infections. The best-studied is gonorrhea, where the male-to-female transmission probability is around 0.5 for a single sexual contact (estimated by tracing the contacts of 37 males)²⁰ and the corresponding female-to-male transmission probability is around 0.2–0.3 for a single exposure (estimated from incidence rates among servicemen consorting with prostitutes)^{21,22}. A study²³ of the steady sexual partners of patients with acute hepatitis B virus (HBV) infections found 20–30 per cent to be infected with HBV, and similarly about 30 per cent of the sexual contacts of syphilis patients have been found to be infected with syphilis²⁴. In both studies, however, it is unclear which person was infected first, so they give only an upper limit to the transmission probability.

Peterman *et al.* also report results for 63 other members of the families of those with transfusion-associated HIV infection, who had no sexual contact with the index case. These 63 comprise 15 wives, 6 husbands, 19 children and 23 others. They lived with the index case for an average of 31 months after the infection was acquired (for a total of 140 person-years of exposure); on average, 27 of these months were before AIDS or HIV infection was diagnosed. None of these 63 other family members showed HIV antibodies. As Peterman *et al.* remark, "When combined with the other published studies, the total

1. May, R.M. & Anderson, R.M. *Nature* **326**, 137 (1987).
2. Peterman, T.A., Stoneburner, R.L., Allen, J.R., Jaffe, H.W. & Curran, J.W. *J. Am. med. Ass.* **259**, 55 (1988).
3. Hyman, J.M. & Stanley, E.A. *Math. Biosci.* (in the press).
4. Castillo-Chavez, C., Cooke, K., Huang, W. & Levin, S.A. *Mathematical Approaches to Ecological and Environmental Problem Solving* (eds Castillo-Chavez, C., Levin, S.A. & Shoemaker, C.) (Springer, New York, in the press).
5. Dietz, K. & Hadler, K.P. *Math. Biosci.* (in the press).
6. Pedersen, C. *et al. Br. Med. J.* **295**, 567 (1987).
7. Goudsmit, J. *et al. Lancet* **ii**, 178 (1986).
8. Lange, J. *et al. Br. Med. J.* **293**, 1459 (1986).
9. Burger, H. *et al. Am. J. Med.* **81**, 5 (1986).
10. Eales, L.J. *et al. Lancet* **i**, 999 (1987).
11. Kreiss, J.K. *et al. Ann. Intern. Med.* **102**, 623 (1985).
12. Jason, J.M. *et al. J. Am. med. Ass.* **255**, 212 (1986).
13. Allain, J.P. *New Eng. J. Med.* **315**, 517 (1986).
14. Smiley, L. *et al. Abstr. 26th Intersci. Conf. Antimicrob. Agents Chemotherapy* (New Orleans, 1986).
15. Padian, N.S. & Winkelstein, W. *Abstr. 26th Intersci. Conf. Antimicrob. Agents Chemotherapy* (New Orleans, 1986).
16. Saltzman, B.R. *Abstr. 26th Intersci. Conf. Antimicrob. Agents Chemotherapy* (New Orleans, 1986).
17. Redfield, R.R. *et al. J. Am. med. Ass.* **253**, 1571 (1985).
18. Fischl, M.A. *et al. J. Am. med. Ass.* **257**, 640 (1987).
19. Mann, J.M. *et al. J. Am. med. Ass.* **256**, 721 (1986).
20. Platt, R., Rice, P.A. & McCormack, W.M. *J. Am. med. Ass.* **250**, 3205 (1983).
21. Holmes, K.K. *et al. Am. J. Epidemiol.* **91**, 170 (1970).
22. Hooper, R.R. *et al. Am. J. Epidemiol.* **108**, 136 (1978).
23. Mosley, J.W. *Am. J. med. Ass.* **270**, 253 (1975).
24. Centers for Disease Control *STD Fact Sheet* Publ. 79-8196 (US Dept Health Educ. Welfare, Atlanta, 1979).
25. Wahn, V. *et al. Lancet* **ii**, 694 (1986).

Robert M. May is Class of 1877 Professor of Zoology at Princeton University, Princeton, New Jersey 08544, USA.

Neural networks

Learning from a computer cat

Alun Anderson

THEORY is making a comeback among neurobiologists, propelled by the potential provided by ever-cheaper computer power. Cynics will say they have seen it all before; and that once again high expectations will be disappointed. But the new generation of neural-network models looks different from the theoretical models of the past: they yield unexpected results, and they make complex phenomena emerge from simple principles in a way that seems reassuringly biological.

Among the most interesting of the new attempts, described at a recent meeting*, to seek analogies between the properties of simulated neural networks and real populations of neurons in the brain are those of Terence Sejnowski and Sidney Lekhy (see below) and of David Zipser and Richard Andersen. The latter, as reported on page 679 of this issue¹, shows just how happily a computer model can, with imagination, be married to data from neurophysiological recording.

Both pairs of researchers use a three-layered neural network model of the kind that is already generating tremendous excitement amongst computer scientists and physicists. How do such neural network models work? And what insights might they provide for neurobiologists?

Shape from shading

Sejnowski and Lekhy set their network the task of computing curvature from shading in an image. The computation is one that our nervous system can easily solve (see, for example, Ramachandran's recent letter to *Nature*²); it is what gives us the ability to gain an impression of depth from changing shades of grey in a picture.

In a neural-network model the outputs and inputs are specified but how the network does the computation to get from input to output is not. An error-correction algorithm, which alters the strengths of connections in the network, is provided to train the network. It is the solution that the network adopts, and its similarity and differences to the behaviour of neurons in biological systems, that is of interest.

The input layer of the network was arranged as an array of 122 overlapping on- and off-centre-surround units, much like those found in the retina and lateral geniculate nucleus of the cat. The layer was 'shown' a sequence of 2,000 images of

parabolic shapes, differing in the direction from which they were illuminated, their degree and direction of curvature and the location of the centre of their surfaces. To read the curvature of the shapes, correlated output was wanted in 24 output layer units, each differently tuned to curvature and orientation. Between them they would signal the principal curvatures and orientations of the shapes, independently of their position or direction of

varying lengths, widths and orientations moved across the receptive field. And, if the units of the output layer are tested in the same way, they appear to behave like a special class of cortical complex cell, the end-stopped bar cell.

The point is that the network model allows a different interpretation of the possible function of these cells; rather than looking for edges they may be extracting shape from shading. That the end-stopped bar cells may be detecting curvature has already received support³, but that they may be extracting three-dimensional shape remains untested. The results thus demonstrate not only that a simple learning rule can produce a network capable of a complex operation, but that analysis of the trained network can suggest what real brains may be doing — and it is not necessarily what classical experiments suggest. Old data may thus be reinterpreted and new experiments suggested.

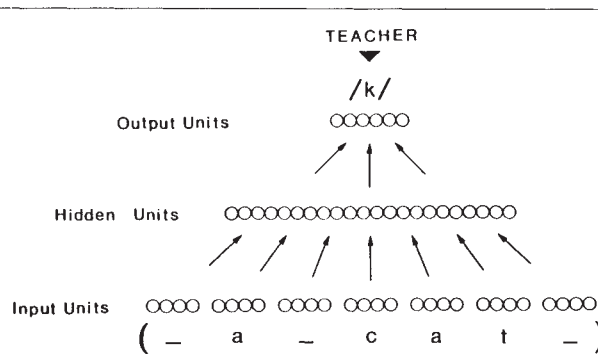
Sounds from words

Sejnowski has used essentially the same type of layered network to look at what happens when other kinds of inputs are turned into feature maps. In collaboration with Charles Rosenberg he taught a network correctly to pronounce English (see figure); the input is words drawn from a 1,000-word dictionary and the desired output the features of spoken English. After 24 hours of training on a Vax computer, the network learns

to generalize and produce the correct pronunciation of words it has never seen before.

Looking at the properties of the hidden units that are doing the work suggests that feature extraction is not a simple hierarchical process. There is a continuum from units that are active most of the time (acting like filters) to those that are rarely active (acting like feature detectors); or from those that are active in many kinds of pattern (part of a distributed network) to those that respond only to certain combinations of features (like a grandmother cell). It seems that some properties are so important they are coded in single cells, others are coding information that is a combination of properties. But they all work together to turn words into speech. Does this 'semi-distributed' representation correctly mirror biological reality? The idea is novel but will not be easy to test. In the computer model, functions of cells can be divined because it is easy to find out where every single cell takes its connection from and sends its connections to. Current neurophysiological techniques can only say where a few cells take their connections from.

Nevertheless, it is still possible to tie



False words — a schematic arrangement of the NETalk system, which is capable of learning to pronounce English text. Three layers of processing units are used: a window of letters in an English text is fed to an array of 203 input units. Information from this level is transformed by an intermediate layer of 80 'hidden' units to produce patterns of activity in 26 output units, whose values determine the output phoneme. From ref. 9.

illumination. The essence of the desired behaviour of the network is thus that of taking a sensory input and turning it into a feature map — exactly what happens in the brain in many stimulus dimensions. To reach the desired performance the network was taught; the difference between desired and actual output was measured and 'back-propagated' through to a 'hidden' layer of 27 units lying between the input and output layers, changing the weights of the connections between units according to a fixed rule (see the discussion in News and Views³). Each unit in each layer connects to all the units in the layer below. At first, as the weights are set at random the output is quite meaningless; but after 40,000 trials the network learns accurately to classify the shapes.

Now comes the surprising part. The network has been trained to extract curvature given an input of activity in cat-type centre-surround receptors. But if the receptive fields of the hidden layer units are investigated using the classical bar and edge stimuli of neurophysiological experimentation, they are found to have acquired properties very like those of the simple cells of visual cortex. Optimal responses can be elicited by bars of

*Society for Neuroscience 17th Annual Meeting, New Orleans, 16-21 November 1987.

together elegantly a network model and neurophysiology, as Zipser and Andersen show in this issue¹. These authors seek to explain how information about the position of an object on the retina is combined with information about the direction in which the eyes are pointing so that the actual location of the object in relation to the body is known — essential information if a hand is to reach out to grasp the object.

In monkeys, lesion studies suggest the two kinds of information are combined in region 7a of the posterior parietal lobe. But neurons from this region do not directly code true spatial position; rather, they seem to respond to a combination of retinal locations and eye positions. In Zipser and Andersen's model, the input layer is composed of an array of units on which retinal and eye positions are dis-coordinately represented; the representations of each are like those recorded electrophysiologically in the monkey brain. The network was trained so that the output layer came to present true spatial location, output errors being back propagated through the system.

After training, the properties of the hidden units were found to match closely those found by electrophysiological recording from neurons in layer 7a. That is to say, the model supports the view that the neurons do not directly code spatial location but are combining retinal and eye position information in a characteristic manner that is a step on the way to a direct reading of spatial position.

The value of the result lies in the possibility that neural networks provide for analysing the computations that the cortex is doing. The obvious worry — that there is no obvious biological process akin to back propagation — may not be serious. Other error-correction algorithms may produce essentially the same result.

Hard questions

Neural-network modelling looks set to help formulate and sharpen theories of what small ensembles of neurons do, but it would be wrong to think it can quickly provide all the answers to the question of how the brain works. Higher-brain function remains a separate domain. Network models of language pronunciation are investigations of mapping; they are not models of how the human brain handles language.

There should be little surprise then that, as P.N. Johnson-Laird recently pointed out in a News and Views article⁵, network models do not capture all the properties of human language learning. For human tasks — like reasoning, thinking about how the brain works, and so on — the rule-based systems of psychologists who study 'mind' will continue to provide more parsimonious explanations. But networks will evolve. At present, they have nothing remotely close to the richness of cortical

Anyone for Tanis?



AN exhibition entitled *The Gold of the Pharaohs* at Edinburgh City Art Centre from 2 February to 30 April brings to Britain for the first time the remarkable treasures from a series of royal graves at Tanis, the 'Thebes of the north', capital of the Pharaohs of the twenty-first and twenty-second dynasties. The site is now a vast mound, 130 km north-east of Cairo and 50 km from the Mediterranean; its tombs were found in a cluster of rock-cut chambers below a temple courtyard. The excavations, carried out by French archaeologist Pierre Montet from 1939 to 1946, produced the richest finds from ancient Egypt since Howard Carter's discovery of the tomb of Tutankhamun in 1922, but the outbreak of the Second World War prevented their achieving the fame they deserved. It will therefore come as some surprise to most visitors to see the wealth on display — gold funerary sandals, bracelets and other jewellery, and, centrepiece of the exhibition, the great gold funerary mask of the Pharaoh Psusennes I, who reigned from about 1040 to 993 BC. The mask and mummy were encased in a silver sarcophagus, inside a stone coffin and an outer coffin of pink granite. In the same tomb was found the solid silver sarcophagus of a previously unknown Pharaoh, Sheshonk II, whose mummy was also decorated with a gold funerary mask and sumptuous jewellery. The material dates from about 1069 to 730 BC — centuries later than Tutankhamun (mid-fourteenth century BC). Of the thousand superb objects recovered in all, only a selection of the smaller items can be transported abroad; already displayed in Paris in 1987, they are on loan from the Cairo Museum.

Paul G. Bahn

Paul G. Bahn is a freelance archaeologist at 429 Anlaby Road, Hull HU3 6QP, UK.

circuitry, but as that richness is added, new properties may, or may not, develop. And even if networks ultimately fail to provide insights into human abilities there are many simpler problems in need of solution along the way. As Sejnowski put it: "frogs don't play chess — but there are plenty of things we don't understand about frogs".

1. Zipser, D. & Andersen, R.A. *Nature* **331**, 679–684 (1988).
2. Ramachandran, V.S. *Nature* **331**, 163–165 (1988).
3. North, G. *Nature* **328**, 107 (1987).
4. Dobbins, A. et al. *Nature* **329**, 438 (1987).
5. Johnson-Laird, P.N. *Nature* **330**, 12 (1987).
6. Linsker, R. *Proc. natn Acad. Sci. U.S.A.* **83**, 7508 (1987).
7. Linsker, R. *Proc. natn Acad. Sci. U.S.A.* **83**, 8390 (1987).
8. Linsker, R. *Proc. natn Acad. Sci. U.S.A.* **83**, 8779 (1987).
9. Sejnowski, T. & Rosenberg, C. *Complex Syst.* **1**, 145–160 (1987).

□ Alun Anderson is Washington editor of *Nature*.

• Not all neural networks require a teacher to acquire interesting properties. At the Neuroscience meeting (see left), Ralph Linsker (IBM Thomas J. Watson Laboratory) explained how feature detectors can emerge from an untrained network, suggesting that in early development of the nervous system, patterned connections may not have to be specified in detail, but can arise naturally from a generic rule.

In Linsker's model⁶⁻⁸, cells are arranged in a stack of layers, each cell's output to the layer above being a weighted average of inputs received from an underlying region of cells in the next layer down. Inputs are 'hebbian': they increase their connection strength if activity in the input cell correlates with activity in the output cell. The blind workings of this rule have subtle effects. Even random input causes feature analysers — contrast- and orientation-selective cells — to emerge in later layers.

Why does this surprising result emerge? A little thought suggests an answer. Each cell receives inputs from many Hebb synapses; each connection will increase (decrease) its strength when it 'agrees' ('disagrees') with the output. An analogy is that of an election where all those who vote for the winning candidate are listened to more carefully in the next election. Thus, groups that vote the same way emerge if there is any structure in (correlations among) their inputs. Structure can come in the the input from the environment.

Where random inputs are provided to the first layer (analogous to the situation of animals born with their eyes still shut) the wiring provides structure. A cell in the second layer receives more input from cells directly beneath it than from those further away. As connection strengths fluctuate and evolve, the set of units that are central to a particular second-layer cell will come to vote as a block; either as an all excitatory or all inhibitory unit. With a structured input coming from the second layer, further layers will add more structure.

The Hebb rule can be described as producing cells whose output varies as much as possible with variation in input from the previous layer of cells so that they signal the maximum amount of information about what is happening there. This 'infomax' principle, as Linsker calls it, can be made rigorous so that it relates input and output variance in a mathematically definable way.

When this principle is put to work in a simulation, topographic maps can be made to emerge that look identical to those found in the cortex. There is thus the possibility that a part of the order seen in the brain emerges as a result of a general information-processing principle. How exactly this principle could be being executed by real neurons remains to be seen. But if true, there are pleasing implications for evolution — generic principles permit re-routing of information without a complete redesign. Alun Anderson

Galactic nuclei

Stellar ejecta from black holes

M.E. Bailey

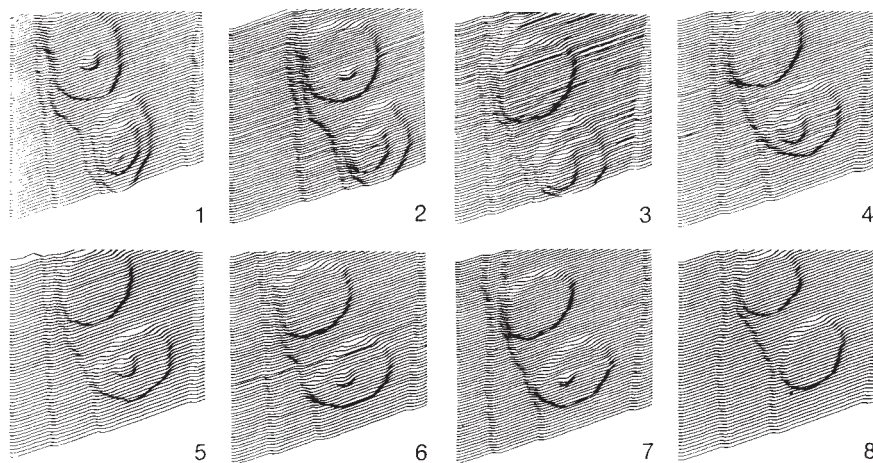
THE resolution of the question of whether there is a massive black hole (at least a million times more massive than the Sun) at the centre of our Galaxy has crucial implications for our understanding of the evolution of galactic nuclei. Although some studies¹⁻⁵ conclude that there is no such central object, others⁶⁻⁸ emphasize that the evidence is inconclusive. The prevailing view, however, is that black holes most naturally explain active galactic nuclei and the quasar phenomenon, and most astrophysicists believe that the black-hole hypothesis gives the best overall explanation of the complex picture now emerging from multi-frequency observations of the galactic centre⁹⁻¹². On page 687 of this issue¹³, Hills gives an extra twist to the problem, showing that a massive black hole in the galactic centre should be a significant source of 'hyper-velocity' stars.

Close binary systems (tightly bound pairs of stars) with orbital velocities V_{orb} greater than the average stellar velocity dispersion are generally expected to survive in galactic nuclei. Such binaries are usually described as being 'hard', and evolve under external stellar perturba-

tions by transferring orbital kinetic energy to the surrounding stars, thus becoming harder and increasingly tightly bound¹⁴. Hills's work shows, however, that if the star cluster contains a massive central object such as a black hole, the close passage of the binary past the massive body could result in a so-called 'exchange' collision. The three-body interaction of the binary with the black hole results in a net perturbation in which one star is captured into a tightly-bound orbit about the black hole, and the other is ejected at high speed into the surrounding galaxy.

The effect is entirely newtonian, and can be approximately understood in terms of the interaction between the two binary components during the period of closest approach to the black hole. (The situation is analogous, in this respect, to the randomizing effects of external stellar perturbations on the orbital energies of binary stars or of comets in the Oort cloud.) Hence, if one component of the binary (of semi-major axis a and total mass about that of the Sun, $1 M_{\odot}$) is regarded as being in an orbit about the black hole (mass M_h) with a distance of closest approach b , its velocity V at this point will

Real-time images of surface diffusion



Scanning tunnelling microscopes (STMs) can be used to observe dynamic processes as they occur on surfaces. R.C. Jaklevic and L. Elie of the Ford Motor Company describe in a recent issue of *Physical Review Letters* (60, 120–123; 1988) how 'footprints' left in a gold surface by deliberate touches of the STM tip slowly fill in over a period of 2 hours. The micrographs above, taken at 16-minute intervals, represent an area $400 \times 400 \text{ \AA}^2$. The step-height is $2.5 \text{ \AA}$ — the height of a single gold atom — so that each footprint represents the loss of about 3,200 atoms. The holes fill in at a rate of only 6–9 atoms per minute, a rate unaffected by the scanning height of the STM tip or the size of tunnelling current used to detect the surface form. The authors tried to observe single atoms 'jumping' by leaving the tip at the edge of a crater, but failed possibly because of the rarity of such events, or because they are too rapid. (Courtesy of R.C. Jaklevic.) □

have a magnitude V of the order of the free-fall speed $(2GM/b)^{1/2}$, where G is the gravitational constant. During the part of the orbit closest to the pericentre, however, the second binary component will perturb the first star by an amount Δv , of the order of their mutual orbital velocity $V_{\text{orb}} \approx (GM/a)^{1/2}$. The change in the orbital energy per unit mass of the first star about the black hole is therefore $\Delta E = (V + \Delta v)^2/2 - V^2/2 \approx V \cdot \Delta v$. As V is generally much greater than V_{orb} , ΔE can be greater than the binding energy per unit mass of the binary and the net effect of the interaction will be to break the binary apart. In this process one star is transferred into a much more tightly bound orbit about the black hole and the other is ejected with a velocity of about $(2V\Delta v)^{1/2}$. For example, with $a = 0.01$ AU, $b = 1$ AU and $M_h = 10^6 M_\odot$, the ejection velocity is about $5,000 \text{ km s}^{-1}$.

In practice, of course, the situation is more complicated than this simple argument suggests; only a relatively small range of impact parameters actually leads to the successful ejection of a whole star. If the binary passes too close to the black hole, for example, individual stars might be directly put asunder; or if b is too large the relative velocity perturbation could be too small to ensure the binary's disruption. Hills has numerically evaluated the resulting ejection velocities for a wide range of initial conditions and parameter values, and plots a quantity that is essentially the average speed of the ejected stars.

In this way, assuming that a reasonable fraction (about 1 per cent) of the stars surrounding the black hole are binaries with suitably small semi-major axes, Hills shows that even a modest black hole in the galactic centre should be a prolific source of hypervelocity stars. Indeed, he argues that a massive black hole in the centre of our Galaxy should eject stars at a rate of around one per 10^4 yr with a speed of about $4,000 \text{ km s}^{-1}$. This rate of production of hypervelocity stars is sufficiently great that one might have expected that we should already have observed at least one such star in the solar neighbourhood, and provides some ground for the hope that the detection of one such star will give the elusive 'smoking gun' long sought as firm evidence for massive black holes in galactic nuclei. Unfortunately, though, the present lack of such evidence cannot convincingly be used as proof of absence.

This work will undoubtedly rekindle interest in the detection of fast-moving stars in the solar neighbourhood (which itself may have important ramifications for understanding the large-scale dynamics of the Galaxy and its massive halo); but if such bodies are not found, their absence will be used as a further argument against there being a massive black hole at the centre of the Galaxy. If this conclusion is sustained, it would have

significant implications for the understanding of active galaxies and of galactic nuclei generally. On the other hand, the proportion of close binaries in galactic nuclei is not well known; it could be less than the 1 per cent assumed by Hills. A detailed study of the formation, evolution and survival of close binaries in and around the dense stellar density cusp about a massive black hole would be an important development from this work.

By suggesting a novel constraint on the mass of a black hole in the galactic centre, Hills not only provides the answer to an intriguing dynamical question but may also trigger new observational and theoretical developments. The underlying question of the nature of energetic nuclear activity in galaxies remains open, but the

mass and even existence of the supposed black hole at the galactic centre may have to be reconsidered. $\square$

1. Ozernoy, L.M. *Observatory* **96**, 67–69 (1976).
2. Bailey, M.E. & Clube, S.V.M. *Nature* **275**, 278–282 (1978).
3. Bailey, M.E. *Mon. Not. R. astr. Soc.* **190**, 217–225 (1980).
4. Ozernoy, L.M. *Astr. Tsirk.* No.1342, 1–5 (1984).
5. Allen, D.A. & Sanders, R.H. *Nature* **319**, 191–194 (1986).
6. Allen, D.A., Hyland, A.R. & Jones, T.J. *Mon. Not. R. astr. Soc.* **204**, 1145–1152 (1983).
7. Bailey, M.E. & Sparks, W.B. *Mon. Not. R. astr. Soc.* **204**, 53P–58P (1983).
8. Genzel, R. & Townes, C.H. *A. Rev. Astr. Astrophys.* **25**, 377–423 (1987).
9. Oort, J.H. *A. Rev. Astr. Astrophys.* **15**, 295–362 (1977).
10. Lacy, J.H. *et al. Astrophys. J.* **241**, 132–146 (1980).
11. Lo, K.Y. *et al. Nature* **315**, 124–126 (1985).
12. Crawford, M.K. *et al. Nature* **315**, 467–470 (1985).
13. Hills, J.G. *Nature* **331**, 687–689 (1988).
14. Lightman, A.P. & Shapiro, S.L. *Rev. Mod. Phys.* **50**, 437–481 (1978).

M.E. Bailey is in the Department of Astronomy, University of Manchester, Manchester M13 9PL, UK.

Palaeontology

Several steps forward for Eocene primates

R.D. Martin

RECONSTRUCTIONS of likely phylogenetic relationships among living primates have now established fairly convincingly that there was an early division between strepsirhines (lemurs and lorises) and haplorhines (tarsiers, monkeys, apes and man). The relationships of the earliest known fossil primates to these two main groups of living primates are, however, subject to continuing uncertainty. The detailed analysis of the ankle and wrist of very early adapid primates by Beard *et al.*, described on page 712 of this issue¹, leads to the identification of features that may at last indicate a definite relationship between these primates and strepsirhines.

The Plesiadapiformes, known mainly from Palaeocene deposits, are only distantly related to later primates (if at all) and it has become common to refer to them as 'archaic primates' to emphasize their distinctiveness. Fossil evidence of undoubted 'primates of modern aspect'

dates back to the early Eocene, at which time they appear fairly abruptly in the fossil record. During the Eocene epoch (38–55 million years ago), the world's climate was considerably warmer and tropical forests extended much farther northwards than today. Accordingly, primates — now mainly restricted to tropical and subtropical regions of the southern continents — are relatively well documented from the Northern Hemisphere of the Eocene.

Reliably identified Eocene primates from both North America and Europe are usually grouped into two families: the Adapidae and the Omomyidae. Perhaps because of a general tendency to seek direct ancestors in the fossil record, several authors have linked the Adapidae (lemuroids) to the strepsirhine primates and the Omomyidae (tarsioids) to the haplorhine primates. Until recently, however, there has been little evidence to

indicate that the Adapidae are specifically related to modern lemurs and lorises, and suggestions of a direct relationship between the Omomyidae and tarsiers have come increasingly under fire^{2–4}. The work of Beard and collaborators provides a welcome clarification.

Reconstruction of the phylogenetic relationships of the strepsirhine primates is greatly hindered by a very poor fossil record. There is no fossil record for the lemurs at all in Madagascar

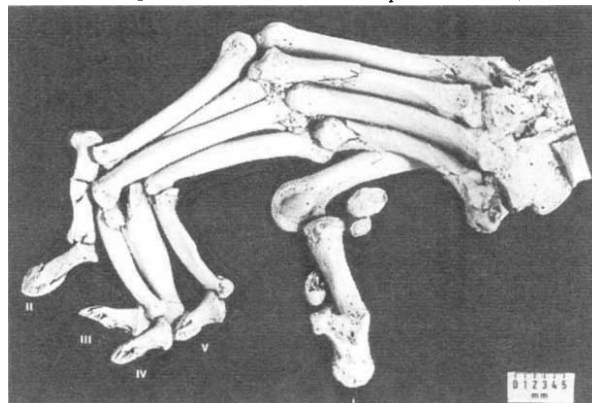


Fig. 1 Left foot of the first adapid specimen from Messel, with flattened terminal phalanges indicating the presence of flat nails as in modern primates. (Courtesy of W. von Koenigswald.)

and the lorises are first documented in Miocene deposits of Africa and Asia, presumably considerably postdating the origin of the strepsirhines. Modern lemurs and lorises do, however, share one striking dental feature — the tooth-comb (or some obvious derivative thereof) in the lower jaw. This apparent shared derived feature, in combination with other evidence, indicates the existence of a specific common ancestor for lemurs and lorises. The Eocene adapids, however, lack a toothcomb and cannot be linked to the strepsirhines by any obvious dental features. Cranial and dental features of the adapids fit the pattern expected for early primates, but they do not indicate a specific relationship with any particular group of living primates.

Until recently, the best-documented adapids were the North American genera *Notharctus* and *Smilodectes*. Abundant associated cranial and postcranial material has been known for some time and *Notharctus*, in particular, has remained one of the best-known fossil primates since Gregory's classic monograph⁵. By contrast, the European forms were for many years documented only by skulls (mainly from *Adapis* and *Leptadapis*) and by isolated postcranial bones attributed to them. The distinctive postcranial morphology of *Adapis* and *Leptadapis* was first described in detail a few years ago by Dagosto⁶ on the basis of individual bones contained in different collections. Since then, a complete articulated wrist attributed to *Adapis* has been described⁷ and new postcranial elements of North American specimens of *Smilodectes* and *Cantius*⁸ have become available. The overall comparison conducted by Beard *et al.*¹ in this issue reveals that in the ankle region (tarsus) there are two features of the talus and a peculiarity of the articulation between the navicular and the cuboid that seem to link the Adapidae specifically to lemurs and lorises. Further, their analysis of the wrist (carpus) suggests that *Adapis* is closer to lemurs and lorises than the North American genus *Smilodectes*. In fact, the detailed morphology of the carpus also indicates that modern lemurs and lorises shared a specific common ancestry after the adapids branched away. This new evidence therefore reinforces the evidence from the tooth-comb that lemurs and lorises constitute a monophyletic group distinct from all other known primates.

There has been another recent, significant advance in the discovery of new articulated postcranial specimens of Eocene primates at the fossil site of Messel bei Darmstadt in the Federal

Republic of Germany (see figures). The oil shale deposits at this middle Eocene site are unusual in that the specimens are remarkably well preserved. Vertebrate skeletons are usually found articulated (and virtually complete) and traces of soft parts are often present. As yet, no complete articulated primate specimen has been found; but three partial skeletons found in the past few years have already been described⁹⁻¹¹ and a fourth is now being studied. Unfortunately, only one of these specimens is associated with craniodental remains, permitting its identification as *Euopolemur*¹¹. The other specimens, which exhibit numerous primate-like features, can be identified as adapids only on grounds of their size. The



Fig. 2 Isolated postcranial elements of the third adapid specimen from Messel. This specimen, like the first, shows the presence of an unusually large baculum (os penis) (bottom left). (Courtesy of W. von Koenigswald.)

striking thing, however, is that in postcranial features all specimens resemble the North American material from *Cantius*, *Notharctus* and *Smilodectes* far more closely than that from the European adapids *Adapis* and *Leptadapis*. Dagosto⁶ showed quite clearly that the postcranial bones attributed to *Adapis* and *Leptadapis* exhibit many peculiarities in comparison to those of *Notharctus* and *Smilodectes*, indicating a very different locomotor pattern. In most postcranial features, *Notharctus* and *Smilodectes* are closer to modern lemurs and lorises, and it seems that *Adapis* and *Leptadapis* became uniquely specialized after their divergence from the common adapid stock. The new Eocene adapid specimens from Messel are therefore of particular importance because they document the existence of European adapids that were considerably closer in their overall postcranial morphology to modern lemurs and lorises. They also indicate that a major division among the European adapids had already taken place by the middle of the Eocene.

Quite apart from the primate specimens found there, the Messel site is of wider importance because of the unique possibilities it offers for detailed reconstruction of an Eocene forest community. For example, more than 50 virtually complete bat skeletons have been recovered and insect remains in their residual stomach

contents have permitted reconstruction of their diets. Complete skeletons of many other mammals have been recovered, including an apatemyid which possessed a specialized anterior dentition and attenuated fingers, convergently resembling the adaptations of the aye-aye (*Daubentonia*) for feeding on wood-boring larvae¹². The remarkable degree of preservation of the mammalian specimens obtained from Messel is shown by the fact that for *Propalaeotherium* (an early member of the horse group), three pregnant females have been recovered, showing that a single, well-developed fetus was typical¹³.

Elation arising from the new discoveries at the Messel site has been tempered by the disturbing news that the local government planned to use it as a regional rubbish dump. Excavation of the fossil deposits is currently being carried out by scientists from three German museums and, with luck, a complete primate specimen will soon be discovered. In view of the remarkable discoveries that have already been made, and given the unusually good preservation of the specimens, plans for rubbish disposal at the site should be completely abandoned. After protests to the local government in Wiesbaden, the plans for rubbish disposal were partially suspended¹⁴. The latest news is that the president of the local government will personally press for the preservation of the Messel deposits in their entirety for palaeontological investigation. It is to be hoped that he will be successful, so we can learn more about the biology of the European primates of 50 million years ago. □

1. Beard, K.C., Dagosto, M., Gebo, D.L. & Godinot, M. *Nature* **331**, 712–714 (1988).
2. Cartmill, M. in *Evolutionary Biology of the New World Monkeys and Continental Drift* (eds Cocchion, R.L. & Chiarelli, A.B.) 47–65 (Plenum, New York, 1980).
3. Schmid, P. in *Primate Evolutionary Biology* (eds Chiarelli, A.B. & Corruccini, R.S.) 6–13 (Springer, Berlin, 1981).
4. Aiello, L.C. in *Major Topics in Primate and Human Evolution* (eds Wood, B.A., Martin, L.B. & Andrews, P.J.) 47–65 (Cambridge University Press, 1986).
5. Gregory, W.K. *Mem. Am. Mus. Nat. Hist.* **3**, 49–243 (1920).
6. Dagosto, M. *Folia primatol.* **41**, 49–101 (1983).
7. Godinot, M. & Jouffroy, F.K. in *Actes du Symposium Paléontologique Georges Cuvier* (eds Buffetaut, E., Mazin, J.M. & Salmon, E.) 221–242 (Montbéliard, Paris, 1984).
8. Rose, K.D. & Walker, A.C. *Am. J. phys. Anthropol.* **66**, 73–89 (1985).
9. von Koenigswald, W. *Paläont. Z.* **53**, 63–76 (1979).
10. von Koenigswald, W. *Carolinea* **42**, 145–148 (1985).
11. Franzen, J.L. *Cour. Forsch.-Inst. Senckenberg* **91**, 151–187 (1987).
12. von Koenigswald, W. & Schierning, H.-P. *Nature* **326**, 595–597 (1987).
13. Hell, R., von Koenigswald, W., Lippmann, H.G., Graner, D. & Heunisch, C. *Fossilien der Messel-Formation* (Hessisches Landesmuseum, Darmstadt, 1987).
14. Martin, R.D. *Folia primatol.* **48**, 221–222 (1987).

R.D. Martin is Professor and Director of the Anthropologisches Institut und Museum, Universität Zürich-Irchel, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland.

Genetic code

Enter a new amino acid

Dieter Söll

BIOCHEMICAL processes in nature are characterized by the great variety of different compounds, yet only 20 amino acids take part in ribosome-mediated protein biosynthesis. Many non-standard amino acids, however, are found in a wide variety of proteins. These 'additional' amino acids arise from post-translational modification of polypeptides synthesized by the established mechanism of protein biosynthesis. Recent evidence, however, indicates that the incorporation of selenocysteine into proteins is achieved in an unconventional way. The paper by Leinfelder *et al.* on page 723 of this issue² provides a crucial detail for the mechanism of specific co-translational insertion of this unusual amino acid into the growing polypeptide chain in response to an in-frame UGA codon in the messenger RNA: there is a minor serine transfer (t)RNA species specific for selenocysteine capable of recognizing this UGA codon. Thus, the protein-synthesizing system is able to encode and use a twenty-first amino acid.

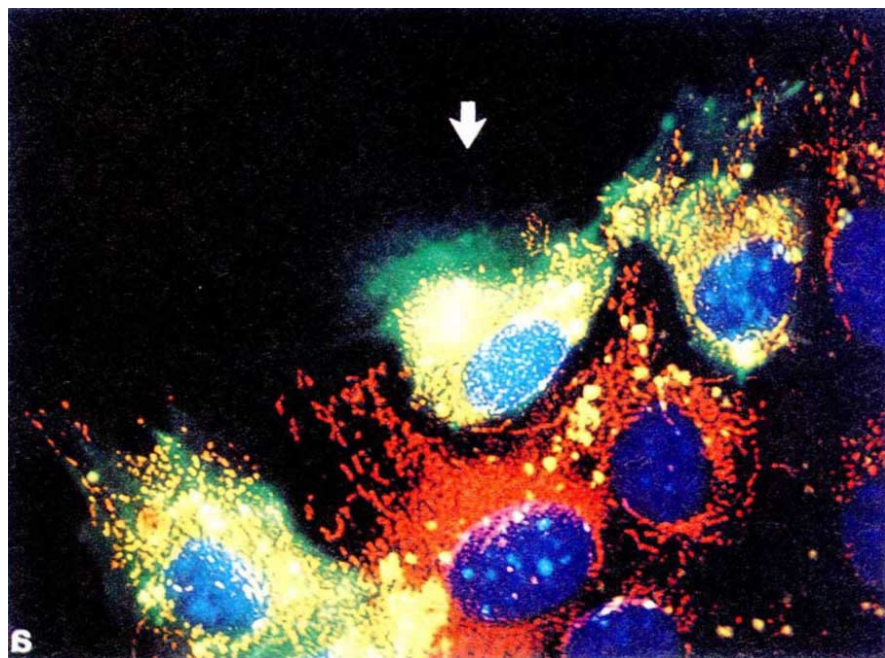
Selenium is a trace element in human diet. There are several enzymes in prokaryotic and eukaryotic cells which contain selenium in the form of selenocysteine in specific positions in the polypeptide chain³. This unusual amino acid is located in the active centre of some selenoenzymes (for example, *Escherichia coli* formate dehydrogenase and mammalian glutathione peroxidase) where the mostly ionized (at neutral pH) selenol group is essential for the redox properties of these proteins³. The redox potential of the selenol group is about 100–150 mV more negative than that of the corresponding thiol group of cysteine. This lower redox potential may be advantageous for the anaerobic conditions in which some of these enzymes are expressed and active.

The importance of selenocysteine is clear from studies of *E. coli* formate dehydrogenase: replacement of this amino acid by cysteine causes a 75 per cent reduction in enzyme activity, while replacement with serine renders the enzyme inactive (A. Böck, personal communication). In both *E. coli* formate dehydrogenase and human and murine glutathione peroxidase, selenocysteine incorporation is directed by an in-frame UGA nonsense codon^{4–6}. *In vitro* mutagenesis of the formate dehydrogenase gene demonstrates that selenocysteine incorporation is co-translational⁷. To study the mechanism of this process, Leinfelder *et al.* collected several mutant *E. coli* strains impeded in

the formation of formate dehydrogenase, and in this way identified four genes required for selenocysteine synthesis and for its incorporation into protein. As their paper in this issue² shows, one of these genes encodes a minor tRNA^{ser} species. They therefore conclude that selenocysteine is formed by modification of serine aminoacylated to a natural suppressor tRNA capable of reading UGA.

These findings raise some interesting questions about protein synthesis: first, how is serine, attached to this tRNA, converted to selenocysteine? A possible intermediate could be the well-known mammalian phosphoseryl-tRNA⁸, which

also reads UGA and has an unknown function. Second, what, in addition to the actual UGA codon, is the signal specifying selenocysteine insertion into a growing polypeptide chain? Are there messenger RNA context effects, novel features of tRNA structure, and/or additional proteins involved? Third, did the genetic code evolve UGA to have some 'extra' coding capacity? This is an especially fascinating question as this codon is used for recognition by natural suppressor tRNAs present in cells^{8,9} as well as its sense use specifying amino acids in certain organelles¹⁰. Fourth, can the mechanism of selenocysteine incorporation elucidate the evolution of the protein-synthesizing system as being flexible enough to accommodate more than 20 amino acids? Indeed, if phosphoserine were co-translationally inserted into proteins it would also use the UGA codon⁸. Fifth, why is there no unique codon for selenocysteine?



ROBBIN DEBIASIO, Gary R. Bright, Lauren A. Ernst, Alan S. Waggoner and D. Lansing Taylor, in a recent issue of the *Journal of Cell Biology* (105, 1613–1622; 1987), demonstrate the combined use of different, specific fluorescent probes to elucidate how separate components of the cell are involved in different functions. Here, Swiss 3T3 fibroblast cells are shown migrating into a wound. Blue, cell nuclei; green, actin; red, mitochondria; yellow, endosomes. In their resting state, these cells do not have any detectable polarity of cell shape or particular distribution of organelles, and organize many of their actin filaments into stress fibres. In this micrograph, cells migrating into a wound are shown to become polarized, with the nucleus at the rear of the cell, the endosomes and mitochondria to the front side of the nucleus facing the direction of migration, and actin-based stress fibres at the rear of the cell. In the early stage of wound healing shown here (3 hours after wounding), the region directly behind the spreading lamellipodia (arrow) displays diffuse fluorescence, indicating that the actin filaments are not well organized. Later in the wound-healing process (not shown here, but see page 1,619 of the paper by DeBiasio *et al.*), both the endosomes and the mitochondria move into the lamellipodia and approach the leading edges of the cells. The cells become elongated and highly motile, and do not have detectable actin stress fibres. This method, combining fluorescence microscopy with the use of selective fluorescent probes, is a powerful approach for elucidation of chemical and structural changes in single, living cells during, for example, endocytosis, cell division and cell motility. □

Does its formation from seryl-tRNA preclude this possibility? If so, it could be that the anticodon of the tRNA is necessary for recognition by seryl-tRNA synthetase.

A final point concerns tRNA. The selenocysteine tRNA gene structure predicts an unusual tRNA sequence (eight base pairs in the acceptor stem) with unusual processing reactions to produce its mature 5' end. Presumably, this structure is needed for recognition by the enzyme(s) converting serine to selenocysteine. Based on our current knowledge of seryl-tRNA synthetase recognition¹¹⁻¹³, it appears that the 5'-terminal three base pairs crucial for synthetase recognition are conserved.

What of the regulation of tRNA gene expression: is this gene expressed only under anaerobic conditions when its product is needed for the synthesis of selenoproteins, or is it present as a general UGA suppressor even in aerobically grown *E. coli*? Why was it not detected as such in previous studies? The conversion of serine to selenocysteine could be another example of modification of the aminoacyl moiety of aminoacyl-tRNA. Like the formylation of methionine attached to the prokaryotic initiator tRNA^{14,15} and the conversion of glutamate to glutamine during the synthesis of glutamyl-tRNA in many organisms¹⁶, it is an important pathway for providing 'additional' amino acids. Examination of selenocysteine incorporation will also shed light on phosphoseryl-tRNA⁸ formation, the only unsolved reaction of this type.

The discovery of the mechanism of selenocysteine incorporation improves our understanding of the importance and diversity of tRNA in biochemical pathways. What other surprises could this versatile molecule hold? □

1. Uy, R. & Wold, F. *Science* **198**, 890-896 (1977).
2. Leinfelder, W., Zehlein, E., Mandrand-Berthelot, M.-A. & Böck, A. *Nature* **331**, 723-725 (1988).
3. Stadtman, T.C. *FASEB J.* **1**, 375-379 (1987).
4. Zinoni, F., Birkmann, A., Stadtman, T.C. & Böck, A. *Proc. natn. Acad. Sci. U.S.A.* **83**, 4650-4654 (1986).
5. Chambers, I. et al. *EMBO J.* **5**, 1221-1227 (1986).
6. Sukenaga, Y., Ishida, K., Takeda, T. & Takagi, K. *Nucleic Acids Res.* **15**, 7178 (1987).
7. Zinoni, F., Birkmann, A., Leinfelder, W. & Böck, A. *Proc. natn. Acad. Sci. U.S.A.* **84**, 3156-3160 (1987).
8. Hatfield, D. *Trends biochem. Sci.* **10**, 201-204 (1985).
9. Geller, A.I. & Rich, A. *Nature* **283**, 41-46 (1980).
10. Breitenberger, C.A. & RajBhandary, U.L. *Trends biochem. Sci.* **10**, 478-483 (1985).
11. Grosjean, H., Nicoghiosian, K., Haumont, E., Söll, D. & Cedergren, R. *Nucleic Acids Res.* **13**, 5697-5706 (1985).
12. Normanly, J., Ogden, R.C., Horvath, S.J. & Abelson, J. *Nature* **321**, 213-219 (1986).
13. McClain, W.H. & Nicholas, H.B. Jr *J. molec. Biol.* **194**, 635-642 (1987).
14. Seong, B.L. & RajBhandary, U.L. *Proc. natn. Acad. Sci. U.S.A.* **84**, 334-338 (1987).
15. Dickerman, H.W., Steers, E. Jr, Redfield, B. & Weissbach, H. J. *biol. Chem.* **242**, 1522-1525 (1967).
16. Schön, A., Kannangara, C.G., Gough, S. & Söll, D. *Nature* **331**, 187-190 (1988).

Dieter Söll is in the Department of Molecular Biophysics and Biochemistry, Yale University, PO Box 6666, New Haven, Connecticut 06511, USA.

Ocean Drilling Program

Milankovitch and monsoons

Leg 117 shipboard scientific party*

THE main objective of Leg 117 was to investigate the origin and variability of the Asian summer monsoon and its influence on depositional patterns and the oxygen-minimum zone (OMZ) in the north-western Indian Ocean. The western Arabian Sea is currently characterized by intense summer monsoonal upwelling that draws nutrient-rich water from below and supports high primary productivity and distinctive planktonic assemblages. The OMZ is caused by the organic deposits becoming oxidized and remineralized as they sink. The planktonic indicators of upwelling conditions, as well as wind-borne (aeolian) detritus from deserts and highlands of adjacent lands, are preserved in the sediments deposited on the Owen Ridge and the Oman margin (Figs 1 and 2). To investigate the evolution of these climatic and oceanographic processes, we recovered more than 4,300 m of sediment at 12 sites. Most of the core recovery was of excellent quality and consisted of continuous sedimentary sections that accumulated rapidly. We were particularly interested in identifying long-term (10⁶ yr) upwelling trends, which might be related to the response of monsoon intensity to uplift of the Himalayas, and shorter-term (10⁴-10⁵ yr) variability that might be caused by changes in Earth's orbit and the seasonal distribution of solar radiation (the Milankovitch mechanism).

We obtained cores from the Oman margin, near the Arabian shore, where there is upwelling, a strong OMZ and organic-rich sediments. Cores from Owen Ridge, far from the shore, towards which the nutrient-rich water spreads from the coast, provide an expanded and continuous record of monsoonal variability of the open ocean during the Neogene (24-2 million years (Myr) ago). We drilled one site (720) on the mid-Indus fan to investigate the relationship between sediment influx and elevation of the Himalayas.

* Co-Chief Scientists: Warren L. Prell (Geology Dept, Brown Univ., Providence) and Nobuaki Niitsuma (Shizuoka Univ., Inst. of Geosci., Shizuoka). ODP Staff Scientist: Kay Emeis (ODP, Texas A&M Univ.). Also: David Anderson (Brown Univ.), Steven C. Clemens (Brown Univ.), Pierre Debrabant (Univ. de Lille), Lawrence A. Krissiek (Ohio State Univ.), David M. Murray (Brown Univ.), Graham P. Weedon (Univ. Cambridge), Werner Ricken (Inst. Geologie, Tübingen), Ahmed Nasser Khalfan al-Thobah (Min. Petroleum and Minerals, Oman), Zaher Khalid Sulaiman al-Sulaiman (Min. Petroleum and Minerals, Oman), Jan Bloemendal (Univ. Rhode Island), Akira Hayashida (Doshisha Univ.), J. Otto R. Hermelin (Univ. Stockholm), Dick Vroon (Vrije Univ.), Catherine Nigrini (La Habra Heights), Stacia A. Spaulding (Univ. Nebraska-Lincoln), Toshiaki Takayama (Kanazawa Univ.), T.F. Pedersen (Univ. Brit. Columbia), Graham B. Shimmiel (Univ. Edinburgh), H. Lo ten Haven (Inst. Petroleum, Julich), Ross Barnes (Rosario Geosci. Assoc., Anacortes), Roman A. Bilak (Calgary, Alberta), Cynthia J. Bray (Cornell Univ.), William H. Busch (Univ. New Orleans), Peter de Menocal (Lamont-Doherty Geological Observ.) and Richard Jarrard (Lamont-Doherty Geological Observ.).

Sites 721, 722 and 731, near the crest of Owen Ridge, provide a composite stratigraphic section more than 900 m thick. These sediments reveal the history of Owen Ridge since the late Oligocene (about 32 Myr ago). Until the middle Miocene (10 Myr ago), they were siliclastic deposits from turbidity currents. From then until the present, the ridge became raised above the turbidity currents and biogenic deposits dominated, often containing opal that is thought to reflect the high biotic productivity caused by upwelling. On the other hand, microfossils

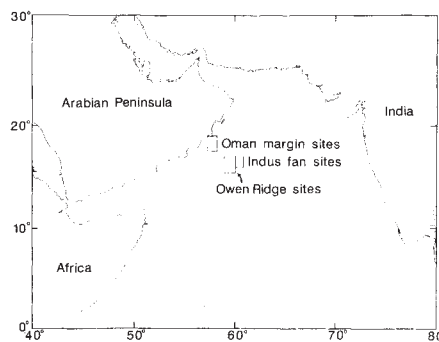


Fig. 1 Location of Leg 117 sites.

indicative of upwelling (various radiolaria and foraminifera) did not appear until the late Miocene. In contrast, both opal deposition and upwelling fauna are found in middle Miocene deposits on the Oman margin. We cannot yet be certain which indicator gives the best measure for the onset of upwelling. The observed time lag between the two locations could represent a seawards expansion of the upwelling environment from the margin.

Patterns of sediment colour, physical properties and magnetic susceptibility of sediments from the late Miocene to the present reveal distinct periodicities (of 400, 100, 41, 23 and 19 thousand years) that match those of the Milankovitch mechanism. These sedimentary cycles seem to reflect variations in primary productivity, preservation of carbon and carbonate, and the proportion of aeolian detritus. Preliminary data show that the magnetic susceptibility varies together with, and probably originates in, the aeolian detritus (indicated by the abundance of quartz).

We cored eight sites on the Oman margin (723-730) at the outer shelf overlying shallow basement, in sediment-filled slope basins, and on or near mid-slope basement ridges (Fig. 2). The range of water depths (300-1,500 m) is that at which the OMZ now impinges on the continental slope. The wide range of settings provides valuable information on the

subsidence of sedimentary basins, the variation of monsoonal upwelling, and the temporal and spatial distribution of organic-rich sediments and their relationship to the OMZ. Most of the sediments on the margin consist of marly nanofossil ooze with admixtures of detrital calcite, opal, foraminifera and scattered phosphorite nodules. All sediments on the margin are rich in organic carbon, with concentrations of 1–5 per cent.

Despite the abundance of organic carbon, sulphate persists in interstitial water to sub-bottom depths often exceeding 100 m, even in rapidly accumulating sediments (faster than 100 m Myr⁻¹). Normally, bacteria use sulphates as a source of O₂ in oxygen-depleted waters, so this finding is unexpected, particularly in sediments within the depth range of the OMZ. A tentative interpretation is that the composition of the organic matter inhibits full use by sulphate-reducing bacteria. Our data still do not resolve the functional relationships between OMZ, sulphate-reduction profile, and high organic-carbon content.

The interaction between the tectonic evolution of the margin and slope basins and the deposition and preservation of

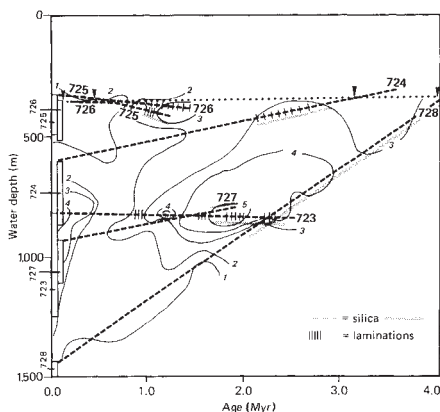


Fig. 3 Age (in Myr) versus depth (m) of Oman margin sites 723–728. Sediment sections recovered at individual sites are plotted on the left according to their present-day water depth and the total depth drilled. The estimated depth of the sediment–water interface through time is indicated by the dashed line for each site and was reconstructed using the presence of the benthic foraminifer *Ammonia beccarii*, which indicates a depth shallower than 350 m. Arrows on the backtracking curves, transitions from < 350 m to > 350 m (dotted line) for sites 728, 724, 725 and 726 are based on this indicator. Isopleths of organic carbon concentrations of the palaeosurface sediments are shown for the correct water depth and age. Darker shading, occurrences of biogenic opal.

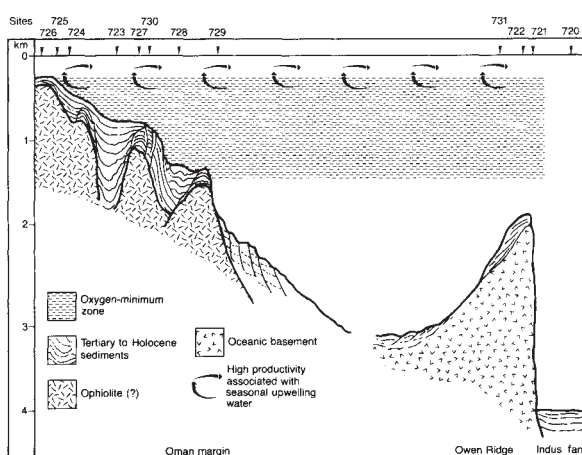


Fig. 2 Leg 117 sites on an idealized depth transect across the continental margin off Oman in the north-west Arabian Sea. Sites 723–730 straddle the depth of a pronounced oxygen-minimum zone that underlies the highly productive surface waters in areas of monsoonal upwelling. Sites 721, 722 and 731 are on Owen Ridge. Hole 720 was drilled on the Indus deep-sea fan to test the influence of Himalayan uplift on long-term evolution of the south-west monsoon.

organic-rich, opal-rich and laminated facies results in both spatial and temporal patterns of deposition (Fig. 3). We constructed subsidence curves of the depositional surface between the level of the modern sea floor and palaeodepths of up to 350 m, inferred at three sites from benthic foraminifer data. The occurrence of siliceous fossils, laminated facies and the percentage of organic carbon are plotted at the appropriate age and depth. Assuming that subsidence occurs at a constant rate, we find that it is 160 m Myr⁻¹ at site 724, 50 m Myr⁻¹ at site 725, 20 m Myr⁻¹ at site 726, and 300 m Myr⁻¹ at site 728. We assume that sites 723 and 727 subsided at a rate of 160 m Myr⁻¹, similar to site 724 in the same sedimentary basin. For site 726, the reconstruction is limited to the past 1.5 Myr, because a hiatus of at least 35 Myr occurred before then.

The data reveal a striking coincidence, especially near 800 m palaeodepth, of organic carbon-rich, opal-rich and laminated facies in the upper Pliocene. These individual sedimentary features also occur in the Pleistocene, though not usually in concert. The late Pliocene mode of deposition and preservation of sub-millimetre laminations may reflect a more intense and stable OMZ. Because the total accumulation rates are not high for this interval, however, the implications for higher or lower rates of upwelling and productivity are unclear. In contrast, sediments of the late Pleistocene and the present-day also contain high organic-carbon content near 800 m water depth, although not as high as the upper Pliocene sediments. They do not contain the laminated, opal-rich facies of the Pliocene sediments. We do not know whether this pattern signifies a change in primary sediment input or a postdepositional change in preservation. □

Daedalus

Inner cleanliness

THOUSANDS of tons of dog-excrement are deposited daily on the parks and pavements of the world's cities. Some regions have 'pooper-scooper' laws compelling the dog-owner to clear up his pet's mess. But Daedalus believes that prevention is better than cure. He points out that excrement is combustible: the heat evolved by burning its organic content is more than sufficient to vaporize its water-content. His ingenious scheme is to burn the excrement while it is still inside the dog.

Daedalus's 'Afterburner' exploits two modern inventions: the catalytic surfaces used in self-cleaning cookers, which catalyse the oxidation of organic matter at fairly modest temperatures; and the nitinol alloys which change shape reversibly on heating. Put in a flame, a piece of nitinol will heat up, bend sharply out of it, cool down, bend back into it, and so on — a simple thermal oscillator. So the Afterburner is designed to sit in the dog's colon, burning the contents catalytically and oscillating rhythmically as it transfers the combustion-heat outwards. The oscillation has two effects. First, it edges the Afterburner peristaltically up the colon, countering the downward flow of fuel; second, it expands and contracts the combustion-chamber, so as to suck in a fresh charge of air and fuel at each stroke. The Afterburner 'breathes' through a flexible pipe which extends to the dog's anus, where its flanged nozzle protrudes and is retained by the reflex sphincter contraction.

The Afterburner has to be inserted by a vet, far enough up the colon not to be excreted itself, but anchored by its tail-pipe so that it cannot travel further. Its pyrotechnic starting-charge is ignited by a fuse, giving a brief burst of smoke from the nozzle; thereafter a soft rhythmic puffing confirms that all is well. Like that of a liquid-fuel rocket motor, the combustion-chamber of the Afterburner is cooled by its incoming fuel. This evaporates some of its water-content before combustion, and also leaks the heat into the dog at a safe and comfortable temperature. The animal will enjoy a higher thermal efficiency, so that its appetite will usefully decrease. A little sterile inorganic ash, puffed down the tail-pipe to a trap in the exit nozzle, may need to be cleared by the owner from time to time.

Sadly, the Afterburner does not make possible the total indoor dog. The creature will still need to urinate, and the calorific value of urine is too low to permit its disposal by combustion. Even so, the urban environment will benefit immensely. Daedalus is now working on a human Afterburner, for use by astronauts, deep-sea divers, and tourists worried about the local sanitary traditions. David Jones

Blood transfusions and HIV vaccinations

SIR—Fuchs¹ and his associates have suggested that large-scale vaccinations against human immunodeficiency virus (HIV) might jeopardize maintenance of adequate blood supplies, especially if such vaccinations occasioned discarding many units of blood with positive HIV-antibody screening tests, or were not completely protective against HIV infection. Testing for HIV-1, screening for type B serum hepatitis, and surrogate testing for evidence of non-A, non-B hepatitis virus infections have already reduced the US national blood supply by 7–14 per cent, because of the need to discard units that test positive. So, how can we maintain adequate, as well as safe, blood supplies?

One simple way is to encourage more volunteer healthy monogamous women who do not use intravenous drugs, along with people over the age of 65, to donate blood 3–4 times annually^{2,3}. Such donors are statistically unlikely to become afflicted with serum hepatitis or retrovirus-associated immune deficiency syndromes unless they have, themselves, received transfusions of blood or blood products^{3,4}. In addition, because blood screening tests for immunodeficiency viruses, lymphocytopathic retroviruses and serum hepatitis viruses are never likely to be perfect, careful prescription of transfusions, along with increasing use of autologous blood and red blood cell savers during surgery will maximize the safety of blood supplies, as well as minimize the transfusional spread of debilitating viral diseases.

JACK W. SHIELDS

Santa Barbara Medical
Foundation Clinic,
Santa Barbara,
California 93102-1200, USA

1. Fuchs, D. *et al.* *Nature* **330**, 702–703 (1987).
2. Shields, J.W. *Lancet* **i**, 850 (1984).
3. Shields, J.W. *Proc. Santa Barbara Med. Fdn Clinic* **3**, 32–40 (1985).
4. AIDS Weekly Surveillance Report (United States) **28** December (1987).

Antarctic earthquakes

SIR—In discussing the low level of seismicity beneath continental ice sheets, Johnston¹ gives the impression that no earthquakes have been located in continental Antarctica. It is true that the level of seismicity is extremely low, but his Fig. 1b omits two earthquakes confirmed in Antarctica, reported in both the *Bulletin of the International Seismological Centre (ISC)* and the centre's *Regional Catalogue of Earthquakes*, and in special studies^{2,3}. The origin parameters are: (1) 4 November 1982, 00:14:19.3 UTC (Universal Time Coordinated), 80.79° S, 36.9° E, depth 0 km (assumed), m_b 4.5; (2) 20 September 1983, 03:27:15.5 UTC, 68.87° S, 122.1° E, depth 0 km (assumed), m_b 4.4.

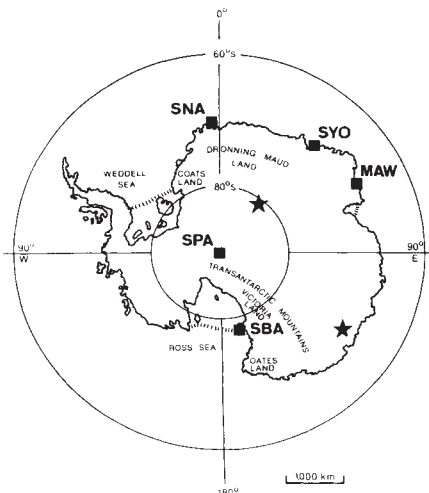
These earthquakes were found by the ISC's 'search' procedure¹, which seeks new events from readings unassociated with earthquakes located by other agencies. Both events were recorded by seismograph stations in Antarctica and by high-gain stations elsewhere. The figure shows the positions of the two events (stars) and the Antarctic stations (squares) reporting the event of 4 November 1982. The 1983 earthquake location is ~200 km inland and ~400 km east of the double earthquake reported on 19 May 1984, shown by Johnston¹. The earthquake of 4 November 1982 happened ~1,200 km from the coast and remains the only confirmed earthquake in the true interior of the Antarctic continent.

The existence of these events in no way negates Johnston's arguments concerning the cause of the low level of Antarctic seismicity, but it cannot be said that the interior of the continent is completely aseismic.

R.D. ADAMS

International Seismological Centre,
Pipers Lane, Thatcham,
Newbury RG13 4NS, UK

1. Johnston, A.C. *Nature* **330**, 467–469 (1987).
2. Adams, R.D., Hughes, A.A. & Zhang, B.M. *Geophys. J. R. astr. Soc.* **81**, 489–492 (1985).
3. Adams, R.D. & Akoto, A.M. *J. Geodynam.* **6**, 263 (1986).
4. Adams, R.D., Hughes, A.A. & McGregor, D.M. *Phys. Earth planet. Inter.* **30**, 85–93 (1982).



Positions of Antarctic earthquakes of 4 November 1982 (upper star) and 20 September 1983 (lower star). Antarctic stations recording the earthquakes of 1982 (squares) are identified by their three-letter international codes.

Not just cachectin involved in toxic shock

SIR—In the issue of 17–23 December 1987, Tracey *et al.*¹ report elegant experiments in baboons showing that neutralization of cachectin (tumour necrosis factor or TNF) by the administration of monoclonal antibodies can prevent the lethal effect of Gram-negative bacterial sepsis. The disease manifestations of such sepsis have long been

recognized as resulting from the various biological effects of lipopolysaccharide (LPS), a component present in the outer membranes of all Gram-negative bacteria. Therefore, the reported findings are in line with earlier work of the same investigators², showing that the toxic and lethal effects of LPS can be prevented by prior administration of polyclonal antibody against cachectin.

As rightly pointed out by Old in his News and Views piece³, cachectin is but one of the many cytokines induced by LPS. The observation that anti-cachectin blocks LPS-induced pathology does not exclude the critical participation of one or more of these other cytokines. One candidate in particular is interleukin-1, a cytokine which not only shares with cachectin all known biological effects, but is also produced by the same cells and under the same circumstances as cachectin. Another candidate is interferon- γ (IFN- γ), a T-lymphocyte-derived cytokine which is known to potentiate both the production⁴ and the actions^{5–7} of cachectin.

That T lymphocytes and their lymphokines are involved in LPS-induced effects is evident from the recent demonstration that administration of anti-IFN- γ antibodies can profoundly modify the local inflammatory effects of LPS⁸. More significantly, it has also been found that the generalized toxicity and lethal effects of LPS can be prevented by monoclonal antibody to IFN- γ ⁹, in much the same way as reported for anti-cachectin. The system used in this study was a generalized and lethal Schwartzman-type reaction elicited in mice by administration of two consecutive, relatively low doses of LPS.

The known synergistic relationship between IFN- γ and cachectin is such that one might compare LPS-induced pathology to an engine running on cachectin as the fuel and operated by IFN- γ serving as the mechanic activating both the fuel injection command (production of cachectin) and the gear shift (action of cachectin). For the therapy of LPS-associated pathology we can indeed empty the fuel tank, but we might also simply talk to the mechanic.

A. BILLIAU

Rega Instituut, Katholieke Universiteit
Leuven,
Minderbroedersstraat 10,
B-3000 Leuven,
Belgium

1. Tracey, K.J. *et al.* *Nature* **330**, 662–664 (1987).
2. Beutler, B., Milsark, I.W. & Cerami, A. *Science* **229**, 869–871 (1985).
3. Old, L.J. *Nature* **330**, 602–603 (1987).
4. Nedwin, G.E. *et al.* *J. Immun.* **135**, 2492–2497 (1987).
5. Williamson, T.W. *et al.* *Proc. natn. Acad. Sci. U.S.A.* **80**, 5397–5401 (1983).
6. Brouckaert, P.G.G. *et al.* *Int. J. Cancer* **38**, 763–769 (1986).
7. Tsujimoto, M., Yip, Y.K. & Vilcek, J. *J. Immun.* **136**, 2441–2444 (1986).
8. Heremans, H. *et al.* *J. Immun.* **138**, 4175–4179 (1987).
9. Billiau, A., Heremans, H., Vandekerckhove, F. & Dillen, C. *Eur. J. Immun.* (in the press).

Tay-Sachs carriers and tuberculosis resistance

SIR—Rotter and Diamond¹ discuss the causes of high allelic frequencies of certain genetic diseases in specific human populations. Unfortunately, they perpetuate the myth that carriers of the Tay-Sachs (TS) allele have an increased resistance to tuberculosis (TB).

Three papers are cited²⁻⁴ in support of this claim. The first is an entire book in which only three articles mention the possibility of TB resistance in Tay Sachs carriers. None of these articles supports the hypothesis of heterozygote advantage: "...number of observations (of increased fitness of TS carriers) was insufficient to guarantee statistical significance." (ref. 2, p. 100). "In the case of TS [disease], it is difficult to visualize a genetic interaction that converts half-levels of hex-A into a selective advantage." (p. 287). "The biological resistance of Jews to tuberculosis, if it exists, is only relative and can be overcome by external environmental deprivations." (p. 306).

The other articles quoted^{3,4} by Rotter and Diamond contain data that allow the calculation of TS allelic frequencies only and present no direct data linking TS carriers with tuberculosis. On the basis of these two papers, therefore, no valid statement can be made about the relationship between the TS allele and TB resistance.

To date, the only direct study supporting the heterozygote advantage hypothesis is that of Myrianthopoulos and Aronson⁵ who compared TB deaths of TS carriers versus non-carriers. However, these authors point out that their correlations between TS and TB resistance are two small to be significant and only suggest trends". We have presented⁶ an independent data set that shows no statistically significant difference in the incidence of TB between the TS and control group. Thus, there is no known selective advantage in being a carrier for the Tay-Sachs allele.

B. SPYROPOULOS

Department of Biology,
York University,
4700 Keele Street,
Downsview, Ontario,
M3J-1P3 Canada

DIAMOND REPLIES—Spyropoulos misstates the evidence that Tay-Sachs (TS) carriers must have some selective advantage; and also that resistance to tuberculosis probably contributes to that advantage. The evidence is as follows.

The TS heterozygote frequency in Ashkenazi Jews is 0.03–0.11, far higher than could be maintained by mutation⁴. As drift or the founder effects are unlikely explanations^{4,7}, and as TS homozygotes all die before reproducing, heterozygotes

must have some balancing advantage. The advantage probably resides in lipid storage, because Ashkenazi Jews also have high frequencies of two other, unrelated, autosomal recessive diseases of lipid storage besides Tay-Sachs: Niemann-Pick disease and adult-type Gaucher's disease.

Half of the grandparents of TS homozygotes must be TS heterozygotes. The grandparents show the expected selective advantage, in the form of a greater completed family size resulting from greater survival of their children to adulthood^{7,8}.

Death rates from tuberculosis were much lower for those grandparents who were TS heterozygotes⁴ (1 in 254) than for those who were not TS carriers⁶ (10 in 356); the difference is significant ($P < 0.05$). Spyropoulos failed to observe this difference in her independent Toronto data set because she examined grandparents not of homozygotes but of heterozygotes⁶: only one-quarter of the Toronto grandparents would themselves have been heterozygotes.

The frequencies of TS heterozygotes and of grandparental tuberculosis deaths among US Ashkenazi Jews vary in parallel with their geographical origin: both are highest for Jews from Austria, Hungary and Czechoslovakia⁴. A similar distribution of TS is also evident in the Toronto sample that Spyropoulos analysed⁴. Thus,

TS distribution matches that of the putative selective factor, tuberculosis, as is also true for sickle-cell haemoglobin distribution and falciparum malaria. Why should tuberculosis have selected for TS among Ashkenazi Jews but not among other European peoples? Because Jews, being confined to the tubercular urban areas, were under extreme selective pressure to evolve resistance, whereas urban populations of other ethnic groups were continually maintained by immigration from rural areas.

It remains to be discovered whether Niemann-Pick and Gaucher's disease heterozygotes also are resistant to tuberculosis, and how lipid storage produces that resistance.

JARED M. DIAMOND

Department of Physiology,
UCLA School of Medicine,
10833 Le Conte Avenue,
Los Angeles, California 90024, USA

1. Rotter, J.I. & Diamond, J.M. *Nature* **329**, 289–290 (1987).
2. Goodman, R.M. & Motulsky, A.G. *Genetic Diseases Among Ashkenazi Jews* (Raven, New York, 1979).
3. Petersen, G. *et al.* *Am. J. hum. Genet.* **36**, 177S (1984).
4. Petersen, G. *et al.* *Am. J. hum. Genet.* **35**, 1258–1269 (1983).
5. Myrianthopoulos, N.C. & Aronson, S.M. *Sphingolipids, Sphingolipidoses and Allied Disorders* (Plenum, New York, 1972).
6. Spyropoulos, B.T. *et al.* *Am. J. hum. Genet.* **33**, 375–380 (1981).
7. Neel, J. V. in *Genetic Diseases Among Ashkenazi Jews* (eds Goodman, R. M. & Motulsky, A. G.) (Raven, New York, 1979).
8. Myrianthopoulos, N. & Aronson, S. *Am. J. hum. Genet.* **18**, 313–327 (1966).

Cometary organics

SIR—In a recent letter, Greenberg and Zhao¹ write: "No published laboratory residue spectrum provides a perfect fit to the Halley dust emission, although that of Chyba and Sagan is closer than that of Hoyle and Wickramasinghe". The figure shows both published flux curves^{2,3}, from which the reader can judge the validity or otherwise of the above statement. Anyone having the patience to make measurements from the two panels could further verify that the mean square

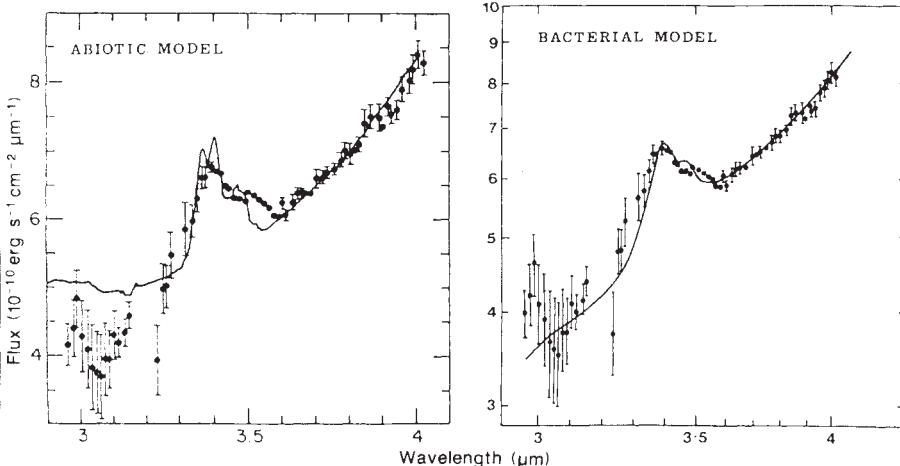
departures from all the observational points for the Chyba-Sagan curve is somewhat more than 2.5 times the corresponding value computed for our model.

F. HOYLE

N.C. WICKRAMASINGHE

Department of Applied Mathematics
and Astronomy,
University College,
Cardiff CF2 4AG, UK

1. Greenberg, J.M. & Zhao, N. *Nature* **331**, 123 (1988).
2. Hoyle, F. & Wickramasinghe, N.C. *Nature* **328**, 117 (1987).
3. Chyba, C. & Sagan, C. *Nature* **330**, 350 (1987).



The curve in the right-hand panel shows our original bacterial model with $T = 320$ K; the curve in the left-hand panel is the multi-component abiotic model of Chyba and Sagan. Points, data for comet Halley on 31 March 1986. (Note that the ordinate scales are different in the two panels.)

Thinking of Bell and Bohr

Michael Redhead

Speakable and Unspeakable in Quantum Mechanics. By J.S. Bell. *Cambridge University Press*: 1987. Pp.212. £25, \$44.50.

Niels Bohr's Philosophy of Physics. By Dugald Murdoch. *Cambridge University Press*: 1987. Pp.294. £32.50, \$59.50.

A LARGE number of books have been published during the past few years on the philosophy of quantum mechanics. They divide into two main classes: popular expositions of the 'paradoxes' and attempted solutions, written generally by physicists, and technical treatises on the philosophical problems involved, written by that new species of academic expert, the philosopher of physics.

Much of the impetus for this renewed interest in the interpretative problems of quantum mechanics stems from the seminal work of John Bell in the 1960s. He challenged the assumptions that lay behind von Neumann's famous proof of the impossibility of hidden variables in quantum theory, augmenting the incomplete description afforded by the quantum-mechanical state vector (or wave function if you prefer wave-mechanical terminology), but showing that any such 'completed' version of quantum mechanics must necessarily involve a feature of nonlocality, of apparently instantaneous action-at-a-distance. This arose from the violation, predicted by the quantum-mechanical formalism, of the famous Bell inequality, which relates certain correlation coefficients in two-particle systems prepared in so-called entangled states, typified by the singlet state of two spin- $\frac{1}{2}$ particles. This inequality was shown by Bell to be a necessary condition that must be satisfied by any theory incorporating local realism.

During the past 15 years many experiments have been carried out to test the predicted violation of the Bell inequality, culminating in the work of Aspect and his collaborators in Paris. They succeeded in showing, relative to making some very plausible assumptions about the working of their apparatus, that the mysterious nonlocal action would have to be effective at space-like separation, that is at superluminal speed of propagation. It is ironic that Bell's work supplements the Einstein-Podolsky-Rosen proof in 1935 of the incompleteness of the quantum-mechanical formalism, predicated on the absence of action-at-a-distance, by showing that a completed quantum mechanics must itself be nonlocal!

It is a great delight to see all of Bell's papers on the foundations and philosophy of quantum mechanics, many of which were published in rather inaccessible sources, collected together and

reprinted by Cambridge University Press. The work of Bell falls into neither of the two classes I mentioned at the beginning. It is serious physics, rather than popularization, but with none of the finicky logic-chopping of the philosophers. Bell always

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Different interpretations — Niels Bohr (top) and John Bell.

works with the simplest examples, and has an amazing talent for putting his finger right on the nerve-centre of any argument.

His view, put very simply, is that the orthodox Copenhagen interpretation, with its talk of a 'cut' between subject and object, its privileged role for classical description on the observer's side of the cut, and what Bell describes as its incomprehensible doctrine of complementarity, should be put aside in favour of the de Broglie-Bohm approach, where instead of particle *or* wave, in complementary situations, we have particle *and* wave. The wave guides the motion of the particle, just like a classical field, and everything is fully objective and well defined at all times.

Bell recognizes that an important element of the Bohrian insight remains, namely that measurement results in general are a product of both 'system' and 'apparatus', rather than revealing some pre-existing property of the 'system'. But we have to bite the bullet and accept the glaringly nonlocal features of this approach. Although the nonlocality involved is 'non-controllable', and cannot be used to send superluminal signals, there are prima facie difficulties with relativity theory. This is a very active area of research among the philosophers of physics. Bell himself poses the problem in the sharpest possible way without really attempting to solve it. He plays with the idea of an aether, picking out a preferred inertial reference frame, but one which is intrinsically unobservable. In a paper entitled, "How to Teach Special Relativity", he emphasizes that this is a perfectly coherent way to understand relativity, following the ideas of Lorentz rather than Einstein.

Bell is generally critical of the Everett-Wheeler 'many worlds' interpretation of quantum mechanics, much espoused these days by cosmologists. He stresses the ambiguity in just when measurements are supposed to take place and the associated splitting of the Universe to occur, and claims that the de Broglie-Bohm approach is a much preferable basis for the application of quantum mechanics to cosmological problems of the whole Universe, where there is in principle nothing 'outside' to play the role of the observer.

By contrast, Murdoch's book attempts a critical assessment of the Bohr interpretation, of which Bell is so dismissive. Murdoch gives a gloss on Bohr's writings, stressing that experiments discover objective values, rather than create them, which he tries to reconcile with the thesis that quantum-mechanical systems do not have intrinsic properties that are independent of measurement. Murdoch emphasizes the influence of pragmatism on Bohr's philosophical thinking, and downgrades the Kantian elements that other writers have discerned in it. He includes discussion of Bell's later work and concludes, on balance, that "there is no outright victor in the Bohr - Einstein contest".

Taken together, the two books give an exemplary introduction to the continuing debates about quantum mechanics. Physicists no longer have any excuse for believing that all questions of interpretation have long been answered, and, as Bell puts it, they could fully understand just how if ever they could spare 20 minutes to think about it! □

Michael Redhead is Professor in the Department of History and Philosophy of Science, University of Cambridge, Free School Lane, Cambridge CB2 3RH, UK.

Associated Press

CERN

The good, the bad and the ugly

Hermann Bondi

Truth and Beauty: Aesthetics and Motivations in Science. By S. Chandrasekhar. University of Chicago Press: 1987. Pp. 170. \$23.95, £19.25.

WHAT a splendid book! Reading it is a joy, and for me, at least, continuing reading it became compulsive.

The fact that this is a collection of lectures (one from 1946, the others bearing dates in the period 1975–1986) means that each of the sections is self-contained, yet there is almost no overlap or repetition. Chandrasekhar is a distinguished astrophysicist and every one of the lectures bears the hallmark of all his work: precision, thoroughness, lucidity. What perhaps I had not foreseen is the depth of the author's historical scholarship. I hasten to add that it was clear to me that if Chandrasekhar was willing to talk about historical or artistic matters, his study would be thorough to perfection; what was not obvious was that so many aspects of these subjects would figure in his lectures.

The last three lectures deal respectively with Milne, Eddington and Karl Schwarzschild, while an earlier group of them concentrates on Shakespeare, Newton and Beethoven. Others are more general. Yet in each case the book is enlivened by reference to events of which the author has personal knowledge, and, for those of us in the know, the discreet references to his own massive contributions are clear and most helpful. For me, born too late to have more than slight personal knowledge of Eddington and of Milne, Chandrasekhar's acute, friendly, yet not uncritical analysis is touchingly effective in helping me see them in the round.

I trust no reader of this review can have missed my enthusiasm for the book, yet I must also add some critical remarks because some aspects of the contents do not fit well into my picture of science. First, to a staunch disciple of Popper like myself, what philosophy of science there is in the book is incomplete, because it does not have any kind of Popperian perspective. Secondly, the very title of the book grates on me (though, fortunately for my enjoyment of it, this turns out to be irrelevant). I am perhaps even more cautious than Popper in assigning truth to science, and also have little faith in the usefulness of beauty as a guiding light. Not only am I bothered by the subjectivity of a judgement of beauty, but it seems to me either to have misled people or to have arisen only well after the original contribution. Thus Chandrasekhar writes so well and so

rightly about the irrelevance to later science of much of the later work of Eddington, of Einstein and of Milne. This work was guided far too much by considerations of beauty, especially mathematical beauty, and far too little by hard, dirty physics.

Again, when Dirac, in his first outstanding contribution, produced the relativistic equation for the electron, its clumsiness and hence ugliness was pronounced. Later work, by Dirac and others, led to new mathematical formulations outstanding in their beauty. Equally, when black holes first arose in the theory, there was nothing particularly elegant about the equations. Much later, Chandrasekhar himself, in his

splendid book *The Mathematical Theory of Black Holes*, put together his own work and that of many others to reveal a pattern of outstanding beauty.

I do not wish to deny that giving earlier discoveries such concise, elegant descriptions is enormously useful for subsequent workers, but evidently it did not contribute to the original discovery. Beauty may be an excellent guide in mathematics, but I doubt its value in physics.

This slightly petulant display of my, perhaps heterodox, views should not stop any potential reader from enjoying this excellent book, as I did so fully. □

Sir Hermann Bondi is Master of Churchill College, Cambridge CB3 0DS, UK.

Buzz-words

Wayne M. Getz

The Biology of the Honey Bee. By Mark Winston. Harvard University Press: 1987. Pp. 281. \$29.95, £23.95.

Books about honey bees are like editions of encyclopaedias. Of both we ask what is new, is the material more readable than in other editions, is the emphasis on a particular branch of knowledge, and at what level is the material presented? In reviewing *The Biology of the Honey Bee*, I will deal with these questions one by one.

The title of the book is apt. Winston presents the full spectrum of basic honey bee biology, in contrast to Seeley in *Honeybee Ecology* (Princeton University Press, 1985) or von Frisch in his exhaustive treatise *The Dance Language and Orientation of Bees* (Harvard University Press, 1967), both of whom concentrated on particular aspects of the subject. The book is intended for biologists who would like to learn more about the honey bee. To quote Bernd Heinrich from the dust jacket, it also provides "... a wonderful reference and entrée to the literature on anything you ever wanted to know about honeybees". For the scientifically well-read non-biologist, a number of the chapters make interesting reading, especially those on worker activities (including nest building, foraging and communication), and reproductive and swarming behaviour. The physiologically orientated chapters, however, which include discussions on the role of juvenile hormone in caste development and the comparative function of the two queen hormones, 9-keto-(E)-2-deconoic and 9-hydroxy-(E)-2-deconoic acids, are unlikely to appeal to a general audience.

Honey bees have been particularly well studied because they are the species of insect that is most economically beneficial to human beings and because their social behaviour (together with that of their ant

relatives) represents a pinnacle of invertebrate evolution. Thus the honey bee is an ideal 'case-study' insect, and Winston's book will serve excellently for students, either as an advanced undergraduate reader or as a basis for a graduate seminar course that leads into the literature (with over 800 references, both the historical and current literature are covered comprehensively). The book should also satisfy professional biologists who would like to see how physiology, behaviour and ecology come together in mediating the organization of a complex insect society, and it can profitably be read by apiculturists wishing to learn more basic bee biology.

The style of writing is clear and relatively informal, making for easy reading in all but the most technical sections. Winston's love for his subject occasionally penetrates through his matter-of-fact style with such phrases as "drones are magnificently constructed for this one task [mating]". There was only one discussion in the book that I had a problem following for lack of clarity. This was the explanation of why workers will not maximize their foraging efficiency by only returning when their crops are full (second paragraph page 180).

If I were asked to fault this book, it would be for the omission of a chapter on apiculture. Winston might argue that his aim was to deal with the basic biology of the bee, and that apicultural aspects can only be treated in detail, as several other authors have already done. However, a 15–20-page chapter on the topic would have rounded off an otherwise complete picture of the honey bee. After all, almost every topic that Winston deals with can itself be greatly expanded, so insufficient depth is no argument for ignoring the symbiotic relationship between honey bees and human beings. Perhaps like a Persian rug, a near-perfect intermediate-level book on honey bees requires a flaw to appease the gods. □

Wayne M. Getz is a Professor in the Department of Entomology, University of California at Berkeley, California 94720, USA.

Back in Athens

Ivor Grattan-Guinness

The Mathematics of Plato's Academy: A New Reconstruction. By D.H. Fowler. Clarendon: 1987. Pp.401. £50, \$98.

THE study of ancient history is beset with special problems. New but reliable sources are not likely to be discovered, and generations of previous historians have already expressed their views, so there seems to be little chance of saying anything which is radically fresh. Yet Fowler's title is no exaggeration: a mathematician by training, he enters into classical scholarship here with a really "new reconstruction" of early Greek mathematics.

The book is divided into three parts. The first, "Interpretations", is largely concerned with his own, which may be summarized as follows. In the so-called 'Euclidean algorithm' (which was known before Euclid) a number a is divided by a smaller number b to produce a quotient with a remainder c ; then c is divided into b and so on. It is well known that the procedure generates a continued fraction expansion of a/b . Fowler takes this fact seriously, and argues that early Greek mathematics was of this kind.

However, whereas the modern subject (which is summarized in the last chapter) draws on the ideas of numbers and number systems common to (many of) our

mathematical theories, these Greek forebears had no such conceptions. For Fowler, they had no abstract (or general) 2, say, but 2 things of some kind or another. In particular, b was divided into a not by any orthodoxly cardinal number of times, but by a 'repetition number' (or a 'gozinter' ('goes into') number, as Fowler pleasantly proposes on p.29); the word 'duple' would be our nearest to an equivalent for the case of 2. Indeed, even talk of division is misleading, because the root of the procedure is the repeated subtraction (called 'anthyphairesis') of b from a until c is obtained.

The first part of the book builds up an interpretation of early Greek mathematics (including Euclid's) from considerations of this kind, partly with the help of specially written Socratic dialogues. Fresh interpretations are suggested for terms such as 'logos' (which means 'calculation' — of some sort! — in mathematical contexts), and for 'rational' and 'irrational' numbers; and a sharp distinction is proposed between the ratio of two numbers (a/b) and the proportion of a quartet of them ($a:b::c:d$).

What evidence can we have for this reconstruction? In the second part of the book Fowler discusses the sources on which any interpretation is based. He draws on the help of several classical scholars and papyrologists, some of whom have reread for him the earliest sources still extant and even have proposed variant readings of details. (Most of the plates tipped into the centre of the book

illustrate some of these texts.) He strongly reinforces the impression — well enough known but of considerable significance here — that much of our knowledge of early Greek mathematics is based on the evidence and opinions of peoples (such as the later Greeks) who lived long afterwards. The third part of the book builds on this theme by discussing later developments, principally those of the later Greeks themselves (by which time the arithmetical thinking mentioned earlier was beginning to emerge, to supplant the earlier doctrines associated with anthyphairesis) and the theory of continued fractions.

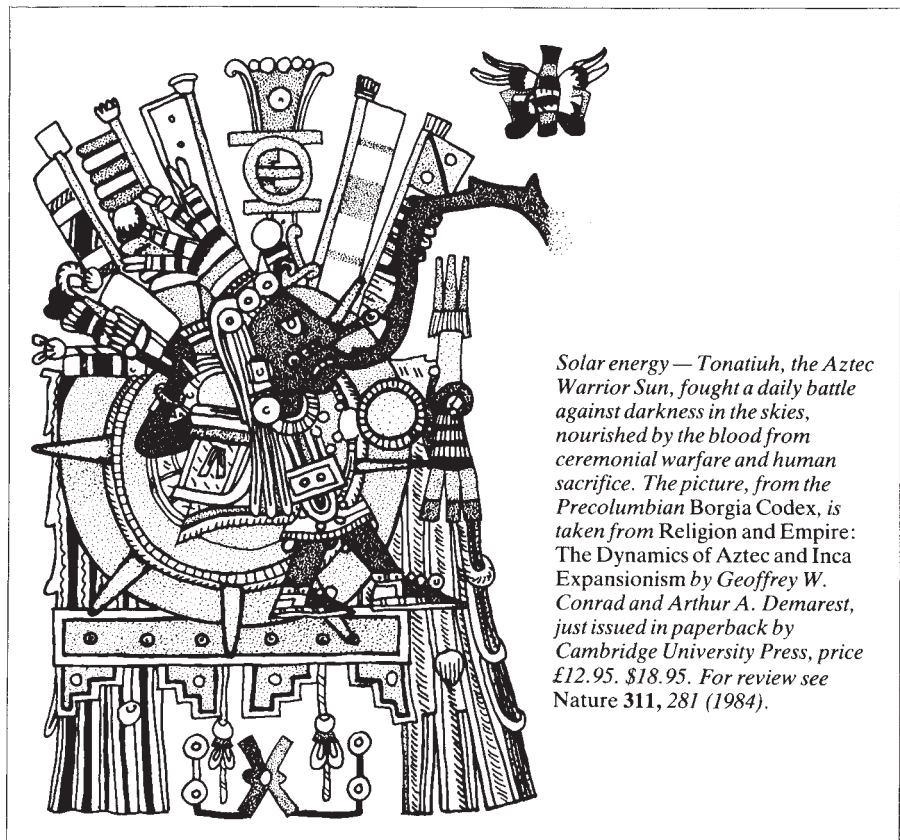
Even from this incomplete summary, it will be clear to the reader that Fowler has dug very deeply into the regions of our presupposed knowledge in framing his new construction. (For example, reading the book provoked in me the thought that our phrasing of fractions as, say, 'five-eighths' would be better said as 'eighths-five'.) He may therefore receive a good measure of hostility and suspicion, especially from those classicists who despise mathematics (p.198), a subject of which they are generally ignorant. Others may see precisely the *merit* of the reconstruction, for we have here genuinely new ideas to work with, and ones which go beyond technical content. Mathematicians should welcome the stress laid upon continued fractions, which is a strangely shadowy branch of mathematics, 'known' but not well known.

For historians there is the thought (not quite stated explicitly in the text) that under this new view they can better conjecture what might have actually been happening in the *Ecole des Hautes Etudes* in north-west Athens all those centuries ago. When Fowler can state difficult theorems and open questions about continued fractions for today, then Plato's veneration of mathematics becomes easier to understand — even though the famous story of the sign over the door of the Academy requiring its entrants to be conversant with geometry is argued to be one of the later accretions to this ancient history! □

Ivor Grattan-Guinness, Middlesex Polytechnic, Enfield, Middlesex EN3 4SF, UK, is President of the British Society for the History of Mathematics.

● Following the completion of Michael Redhead's review (page 667 of this issue) we have received a further book on quantum physics. *The Description of Nature: Niels Bohr and the Philosophy of Quantum Physics* by John Honner is just published by Clarendon, Oxford, and costs £25.

● S. Chandrasekhar, whose new book is reviewed on page 668, has contributed a foreword to Arthur S. Eddington's *The Internal Constitution of the Stars*, first published in 1926 and now reissued in paperback by Cambridge University Press. Price is £12.50, \$19.95. For review see *Nature* 119, 111 (1927).



Solar energy — Tonatiuh, the Aztec Warrior Sun, fought a daily battle against darkness in the skies, nourished by the blood from ceremonial warfare and human sacrifice. The picture, from the Precolumbian Borgia Codex, is taken from *Religion and Empire: The Dynamics of Aztec and Inca Expansionism* by Geoffrey W. Conrad and Arthur A. Demarest, just issued in paperback by Cambridge University Press, price £12.95. \$18.95. For review see *Nature* 311, 281 (1984).

Dr Boyden at the bedside

Vernon Reynolds

Western Civilization in Biological Perspective: Patterns in Biohistory. By Stephen Boyden. Clarendon: 1987. Pp.370. £32.50, \$55.

CAN the ills of our Western way of life be diagnosed? Is our civilization healthy? Has the past produced a vigorous child or a many-headed monster? For Stephen Boyden, our conventional history is inadequate to the task of describing and evaluating modern society; he relies on 'biohistory'. This generates a different kind of understanding, 'biosophy', which is rooted in nature, and sees man evolving from a state of nature to a state of culture. Culture is our human adaptation. In the first place human beings adapted culturally to prevailing (natural) ecological circumstances, but soon undesirable consequences became apparent and we had to make further cultural adaptations. This is a continuing process — today's anti-pollution laws, for instance, are adaptations to unwanted fall-out from earlier cultural adaptations.

This is Boyden's theme, as later chapters show, but the first half of the book consists of biohistory. This is divided into three stages — *primaeval*, early farming and early urban — which lead on to the high-energy stage, the one in which we now live.

Much of what Boyden relates in this book is by no means new. What is new is the synthesis of ideas. It is only at the end, in the appendix, that one realizes what the author is up to. He writes: "My colleagues and I have found check-lists to be an invaluable device for ensuring that . . . all important and relevant variables, concepts, and principles are taken into account in the biohistorical assessment of human situations" (p.325). A series of check-lists follows. "Biohistorical assessment" — in other words, diagnosis of our modern society — is Boyden's aim. He finds it alive but pervasively sick.

In evaluating this enterprise, the first question must be whether a better understanding of our particular kind of modern society can result from study and knowledge of humanity's *primaeval* hunter-gatherer, early farming and early urban past. Those past phases have given rise not only to Western civilization but to all non-Western ones as well.

Toynbee, briefly mentioned by Boyden, argued that each civilization developed because it took up a certain challenge, which set the limits for endeavour for its leaders and its people. Loss of confidence in the value of the challenge led to each civilization's decline and fall. Boyden's pages are not stocked by descriptions of the rise and fall of other civilizations, nor could this one book possibly have taken on more than it has already. But other civilizations have arisen and flourished mightily, and there is much left to be explored in the precise ways that these have developed their particular preoccupations. Spiritual ideas have in the past often been paramount in the minds of the élites controlling civilizations — witness Ancient Egypt — and perhaps the fact that ours is a civilization whose leaders' preoccupations are almost entirely secular may be important in understanding what is especially distinctive about our own day and age.

Darlington, surprisingly, gets no mention. His book *The Evolution of Man and Society*, like Boyden's, tries to understand history and ourselves from an evolutionary perspective. His work differs in approach from Boyden's for it is replete with descriptions and insights into the different trajectories of peoples, cultures and societies all over the world. It has none of Boyden's competence in ecology, none of his awareness of the hazards to health and to the well-being of our own kind of civilization. But it is comparative, and it does try to show how, by taking one

path, or solving one problem, societies and civilizations have succeeded in making certain advances while, simultaneously, sowing the seeds of their own destruction.

Biohistory eschews the fine detail of time and place. Instead we have "evodeviant" cultures, always in danger of worsening life because of the mistakes they make with respect to what is natural for man. This is something of a blanket critique and it runs a very particular danger: what seems to be harmless is accepted as natural, and what is harmful is seen as unnatural. For instance, we are told that the most natural diet — raw fruit and vegetables with a little cooked lean meat — is also the healthiest one. This certainly seems to fit with modern Western dietary ideology, but we need far better evidence than is currently available about the health-related effects of all the varied diets that exist world-wide in order to be sure what is optimal.

What we have in this book is an evolutionary backcloth followed by an ecological critique of Western society. The ideas are distilled and wise, if at times rather bland, and one is left feeling that this is an important strand of modern thinking. It is clear that we are paying a price for our way of life, both individually in terms of psychic health, and ecologically in terms of the continuing damage we inflict on the biosphere.

Boyden ends with a look to the future, in which he concludes that

society must be organized so that the life conditions of people are health-promoting, with regard both to material aspects of experience, such as diet, air quality, and housing, and to behavioural, psychosocial, and intangible aspects which affect, for example, physical fitness, sense of personal involvement, sense of belonging, and sense of self-fulfilment [p.313].

How this might happen is less evident. "Clearly, some new element has to be brought into the adaptive process", something that "transcends national, political, religious, occupational, and racial boundaries" (p.314). This is not an avowal of optimism. Unlike Konrad Lorenz, who after laying out our innate inclinations to go astray was wont to end with a fervent message of hope, Boyden in his final pages reverts to "phylogenetic ineptitudes", "behavioural deficiencies" and the "flaming moth" principle. Certainly he has catalogued Western civilization's mistakes, and his book is a useful reminder to those of us who benefit most from its wealth that the price we are already paying could escalate into a disaster. □

Vernon Reynolds is a Lecturer in the Department of Biological Anthropology, University of Oxford, 58 Banbury Road, Oxford OX2 6QS, UK.

This is a great book,
and an exciting book. . .
readable, worth
reading and
enlightening.

Sir Karl Popper

Cosmology, Physics & Philosophy

by
Prof Ben Gal-Or

A Master Piece!
*Indian Journal of
Physics*

. . .shines a welcome
light in some dark
corners of science.

New Scientist

2nd Edition
Springer-Verlag, 1987
175 5th Ave.,
N.Y., 10010.

Reader Service No.7

A man of universal interests

Andrei Sakharov

The death of Ya. B. Zel'dovich robs physics of one of its guiding lights of the twentieth century. Here Andrei Sakharov reflects on his life and his science, while on p.673 appears a report of two meetings written by Zel'dovich, in collaboration with A. A. Starobinskii, shortly before his death.

ON 2 December, 1987, Academician Yakov Borisovich Zel'dovich died suddenly from a heart attack. He was an outstanding physicist, who made enormous contributions to many branches of science and technology. I am not a specialist in every sphere of his activity (indeed, he was probably unique in the breadth of his interests), and hence I will touch on certain sides of his work only in broad outline. From 1948 to 1968, however, the two of us worked closely together, and I know a great deal of about that period of his life.

In his memoirs, Zel'dovich states that in 1931 (when he was a 17-year-old laboratory technician in the Institute of Processing of Useful Ores) he went on an excursion to the laboratory of chemical physics of the Leningrad Physical-Technical Institute and got into discussion with some of the staff on the forms of crystallization of nitroglycerine. He was invited to work in the laboratory in his free time, and soon afterwards he transferred there officially. So began his path in science, which was to last for 56 years. In 1936 he defended his dissertation for the degree of Candidate of Sciences; later he wrote of the "happy times when permission to defend [a Candidate's dissertation] was granted to people who had no higher education".

In the 12–15 years following his move to the Physical-Technical Institute, Zel'dovich carried out outstanding work on the theory of combustion and detonation, adsorption and catalysis, the fixation of nitrogen, chemical chain reactions and (before, during and after the war) reaction technology. His interest in chemical physics lasted his entire life, and his last work now seems to have been a return to this first love.

The discovery of uranium fission changed Zel'dovich's scientific destiny, as it did that of many other scientists, years before — on a far wider scale — it changed the destinies of us all. His pioneering researches, in collaboration with Yu. B. Khariton, on the theory of explosive and controlled fission chain reactions, were simultaneously the last to be published in the open literature until the veil of secrecy was lifted from the subject. They had a great influence on everyone working in this field. From the very beginning of Soviet work on the atomic (and later the thermonuclear) problem, Zel'dovich was

at the very epicentre of events. His role there was completely exceptional.

In the middle of the 1950s, Zel'dovich found himself a new area of activity — first of all, the theory of elementary particles (in which, soon after the end of the Second World War, there had been a breakthrough, and which has developed up to

of the early Universe; and for his closeness to observations. I shall enumerate in no particular order a few of his contributions to astrophysics and cosmology.

In Zel'dovich's work of 1967, there is a formulation of the problem of the cosmological constant. According to Zel'dovich, it followed from the theory of elementary particles that this constant is small or equal to zero. At the present time, the cosmological constant is one of the central problems in attempts to construct a unified theory of all fields and interactions. In publications with Ya. A. Smorodinskii and S. S. Gershtein, Zel'dovich considered the cosmological limits on the masses of the electron and muon neutrinos. These contributions are examples of the new directions in science that arose during the 1960s, that lie at the junction of cosmology, astrophysics and the theory of elementary particles, and that to a considerable extent are associated with Zel'dovich. Here, the entire Universe plays the role of a gigantic laboratory. Zel'dovich's works on the generation of particles in a gravitational field are of great importance, and include a joint paper with Pitaevskii which contains a remarkable discussion with S. Hawking. The effects of polarization of a vacuum make possible the generation of particles by a classical field. Soon after this discussion, Hawking himself published his famous theory of radiation evaporation of black holes. The effects of polarization of a vacuum (quadratic conformal anomaly), according to the theory of Zel'dovich's colleague, A. A. Starobinskii, lead to cosmological solutions without singularity. (I don't give this formulation my wholehearted support and would propose instead that the main causes of inflation are 'false vacuum' effects.)

Closer to classical astrophysics, but nonetheless important, are Zel'dovich's publications on neutron stars and black holes, and accretion and radiation in solitary objects and binary stars. In his first paper on black holes (1964), he put forward the idea of observing black holes by the radiation from material moving in its gravitational field (simultaneously and independently, a similar proposal was published by Salpeter). Soon afterwards, there followed his work in collaboration with O. Kh. Gusseinov on radiation in

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

Ya. B. Zel'dovich, 1914–1987.

the present time), and then, in the 1960s, the no less dynamic and captivating field of astrophysics and cosmology.

In 1955, Zel'dovich and S. S. Gershtein together put forward the hypothesis of the conservation of the weak charged vector current. This idea (which was discovered independently by Feynman and Gell-Mann) played an important part in the formulation of the theory of weak interactions and the unified theory of weak and electromagnetic interactions, and also in what became known as 'current algebra'. In another paper of the same period, Zel'dovich predicted the existence and certain properties of the Z^0 boson.

For the last 25 years of his life, astrophysics and cosmology had a central place in the thinking of Zel'dovich and his pupils. He was universally acknowledged as a world leader in the field — for the exceptional clarity and concreteness of his physical thinking; for his intellectual daring as a theoretical physicist, which was applied with equal facility to physical laws and theoretical methods, to the formation of Liesegang rings in test-tubes, to the grandiose processes of the explosion of a supernova with the formation of a neutron star or black hole, and to even more extreme processes of the cosmology

Novosti

binary systems, one component of which is a black hole. In a work of 1964 (in collaboration with M.A. Podurets), and in a number of subsequent papers, he considered the dynamics of neutron emission during the formation of black holes. As a result, black holes became accepted as really observable objects, and appropriate observational programmes were worked out and began to be implemented. These programmes, as is well known, have already yielded very interesting results.

Another important aspect of Zel'dovich's interests (in which A.G. Doroshkevich, I.D. Novikov, R.A. Sunyaev, S.F. Shandarin and others also took part) was the formation of galaxies and galaxy clusters. He established the nature of the spectrum on initial perturbations of density, which have observable consequences, and he predicted singularities of the large-scale structure of the Universe — according to that theory, there are gigantic 'black regions' free from galaxies and filled with hot low-density gas of 'pregalactic' composition. There are data suggesting that this is really the case, although the question must still be considered open.

Zel'dovich and his colleagues analysed cosmic electromagnetic radiation and proposed experiments to be carried out in a range suitable for the discovery and investigation of relict radiation. These ideas were not generally known abroad and were not properly exploited by those who did know them. Immediately after the discovery of relict radiation, Zel'dovich recognized it to be of enormous importance, not only as a confirmation of the hot model of the Universe but also as a powerful means of investigating many other important questions in cosmology and astrophysics. In a number of publications (in collaboration with R.A. Sunyaev and others) he examined the effect of various cosmological factors on the anisotropy of relict radiation. As is well-known, this line of investigation has acquired great significance.

In his last years, Zel'dovich published works in which there was an especially profound reflection on the interconnection of the theory of elementary particles and cosmology, considering cosmic domains, and cosmic structures, the astronomical consequences of the rest mass of the neutrino and other postulates of the theory of elementary particles. He attempted to indicate the outlines of what he called 'full cosmological theory' (corresponding to the question of the character of the pre-classical, that is, the quantum-gravitational, stage of development of the Universe, and the origin of the qualitative and quantitative characteristics of the classical stage, including the polytropy spectrum of 'initial perturbations').

All his life Zel'dovich was at the lead-

ing edge of science, and he was always surrounded by people. His effect on his pupils was remarkable; he often discovered in them a capacity for scientific creativity which without him would not have been realized or could have been realized only in part and with great difficulty. An essential factor here was his scientific style and personality — his immense energy, his sensitivity to what was new, his intuition, his striving for theoretical simplicity and elegance, his scientific honesty, and his readiness to admit his own error or acknowledge the priority and correctness of another. In science, Zel'dovich was a humble man (although the manner in which he sometimes took part in discussions, defending what he considered to be the scientifically irrefutable, could give a somewhat different impression). He was almost childishly delighted when he had managed to achieve some important piece of work, or had overcome a methodological difficulty by an elegant method, and felt failures and errors keenly.

To Zel'dovich, it often seemed that he was not professional enough in this or that field, and he summoned up greater efforts to fill in these gaps. Here, too, his approach was creative. He often found new, more comprehensive ways of describing and dealing with a problem. So there came into existence numerous papers and articles of a pedagogic nature, and monographs and books (more than 20 of them) which always included much that was original. His books *Relativistic Astrophysics*, *Theory of Gravitation and the Evolution of Stars* and *The Structure and Evolution of the Universe* (jointly with I.D. Novikov) acquired great renown. It is impossible to overestimate the importance of this side of his work, which helped a great many people to come to science by the most direct route.

Special note should be made of his book *Higher Mathematics for Beginners*. In one of his papers, Zel'dovich wrote: "The so-called 'strict' proofs and definitions are far more complicated than the intuitive approach to derivatives and integrals. As a result, the mathematical ideas necessary for an understanding of physics reach school-pupils too late. It is like serving the salt and pepper, not for lunch, but later — for afternoon tea". I whole-heartedly agree with him on this point.

As I have already said, for many years I worked closely with Zel'dovich. This was a relationship of comradely goodwill, mutual readiness to help and strenuous work towards a common aim. There was no negativism, hostility or signs of unhealthy competition (this, moreover, although the group of I.E. Tamm, to which I belonged, had 'fallen in' to an already well-established team from outside). In the 1950s and 1960s, our offices and homes were next to one another, and

several times a day we would get together to consider a basic theme; we would also talk about general scientific problems. Often we discussed interesting physical and mathematical trifles (what I call 'amateur problems'). Sometimes we played games, as it were, competing in the speed and elegance of solutions (the one who solved it first ran to tell the other at any hour of the day or night).

Much of my own work on fundamental science had its beginning in these contacts. Here is one example. Once Zel'dovich rang me up and said that he had been lecturing to one of the Moscow scientific seminars on his work on the cosmological constant (described above) and had met with puzzlement. I immediately appreciated the importance of the problem, and a few days later I rang him up with a further development, the idea of induced gravitation. Zel'dovich received the idea enthusiastically, and, in his turn, wrote a paper which treated electrodynamics in an analogous way. I, too, understood the role played in the problem of the cosmological constant by the compensation of the boson and fermion contributions — unfortunately, neither of us managed to think the problem through as far as supersymmetry.

I should also like to put it on record that Zel'dovich helped a great number of people in a purely personal way, even in matters of everyday life. Of course, one should not assume that in every case Zel'dovich appeared in the best possible light. He was no angel.

My relations with Zel'dovich were not always unclouded. In the 1970s and 1980s, especially in the Gor'kii phase of my life, hurt feelings and mutual coldness crept in. Zel'dovich strongly disapproved of my social work, which irritated and even frightened him. He once said "People like Hawking are devoted to science. Nothing can distract them". I did not understand why he could not give me the help which, given our relationship, I considered myself justified in asking for. I know that all this tormented Zel'dovich. It tormented me too, as described in my memoirs. Today, the events of those years seem like foam, carried away on the stream of life. Unfortunately, after my return from Gor'kii I met Zel'dovich only once or twice, and then in company, and was hardly able to communicate with him in human terms. Yet another lesson is that it is not always possible to put things off to another day.

Now, when Yakov Borisovich Zel'dovich has departed from us, we, his friends and colleagues in science, understand how much he himself did, and how much he gave to those who had the chance to share his life and work. □

Andrei Sakharov is at the Lebedev Physical Institute, Academy of Sciences of the USSR, Moscow, USSR.

Quantum effects in cosmology

A.A. Starobinskii and Ya.B. Zel'dovich

Two recent symposia on quantum cosmology and quantum gravity* reflected the ardent hope of deriving the ultimate explanation of all of nature from the fundamental laws of microphysics. This goal is more ambitious than that of the classical astronomers, who, given the initial conditions, wished to calculate precisely all the subsequent motions of the planets, stars and so on. The question now is what are the initial conditions for the Universe as a whole that are prescribed by fundamental theory?

These initial conditions refer to the moment of the very beginning of the expansion of the Universe, about 10–15 billion (10^9) years ago. At that moment, nature was characterized by ultracurved space-time, ultrastrong fields and very massive particles not yet discovered in laboratories. Quantum cosmology is the application of quantum theory to this exotic situation. Its renaissance during the past few years was caused mainly by the success of the inflationary model of the Universe (stressed by A.H. Guth, Harvard-Smithsonian Center for Astrophysics).

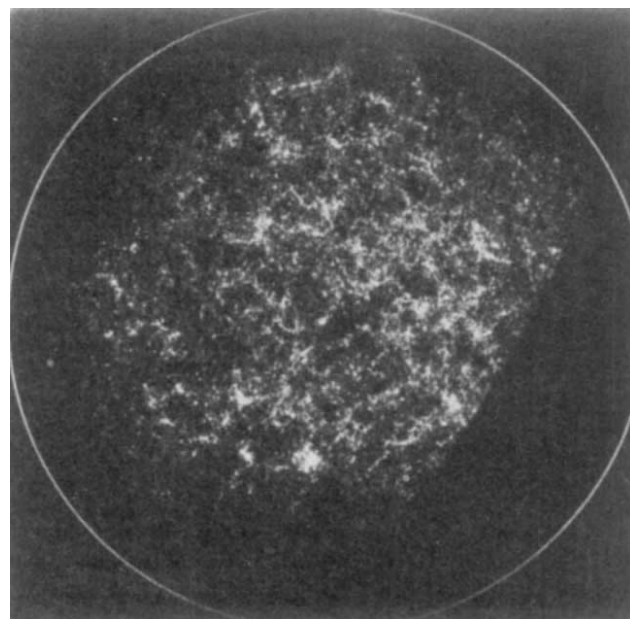
Inflation, a stage of rapid accelerated expansion of the Universe with a scale factor that depends exponentially on time, makes it possible to relate the initial conditions of the Universe to such directly measurable quantities as the anisotropy of the cosmic background radiation temperature, the large-scale distribution of galaxies and the spectrum of stochastic gravitational-wave background. As a result, quantum cosmology seems to be strongly connected with both microphysics and observational astronomy. The most interesting reports discussed the structure of the Universe at very large scales (far beyond the present-day cosmological horizon), the predictability and interpretation of the wavefunction in quantum gravity and quantum cosmology, and the evolution of multi-dimensional ('Kaluza-Klein') models with lagrangians (equations of motion) modified by quantum corrections following from the more fundamental superstring theory.

It is certain that only statistical predictions for the initial conditions of the Universe can be obtained. But a look at the night sky or at a map of distant galaxies shows that this is sufficient: galaxies are distributed without any accurate design (see figure). They do like to be together, however, forming clumps, filaments (with galaxies attached like pearls on a necklace) or surfaces surrounding voids.

The evolution of the Universe since it

was about 0.1 s old is characterized by particle energies available in existing accelerators. The predictions concerning nuclear reactions in the famous 'first three minutes' are confirmed by observations. This is a triumph of the hot Big Bang picture. But the observed abundances of D, ^3He , ^4He and other elements imply that the density of matter consisting of protons and neutrons (such as stars and dust) is less than 10 per cent of the total matter density of the Universe as deduced from the dynamics of clusters of galaxies. There

Descriptions of the very early stage of the evolution of the Universe also combine the well established and still unknown. The idea of inflation now seems inescapable. One needs a period when the matter filling the Universe had very unusual properties — positive energy density ϵ but negative pressure P with $P \approx -\epsilon$. The rigorous description of this matter requires the introduction of an effective (not necessarily fundamental) classical scalar field ψ with a proper potential $V(\psi)$. One of the most important



A map of the 1.27 million galaxies brighter than magnitude 19 in the northern sky. It was prepared by M. Seldner, B. Siebers, E.J. Groth and P.J.E. Peebles (Astr. J. 82, 249–256; 1977) from the Lick Catalog of galaxies drawn up by C.D. Shane and C.A. Wirtanen. The map is broken down into 10 min $\times$ 10 min cells each of which is assigned a brightness proportional to the number of galaxies it contains.

Astronomical Journal

are several arguments that suggest that more than 90 per cent of the total density is in the form of 'dark' matter. This must be some kind of matter that does not interact with the electromagnetic field (so consisting of neutral particles and/or fields) and affected only weakly by forces other than gravity.

'Cosmic strings', also hypothesized as a part of the zoo inhabiting the Universe, would contribute to the total density, but, more importantly, they would generate a significant intensity of gravitational radiation, creating specific perturbations that lead to galaxy formation. If the strings are superconducting as predicted by some variants of fundamental theory (E. Witten, Princeton University), they are also capable of generating very-low-frequency electromagnetic waves that cannot be observed directly but could heat surrounding plasma. The overall picture of the Universe and of its immediate past will remain incomplete as long as the nature of dark matter and the question of strings are not solved.

achievements of this idea is the explanation (Ya.B.Z.) of the very expansion of the Universe. In Einstein's general relativity, gravitational acceleration depends on $\rho + 3P/c^2$ (c is the speed of light) not simply on the mass density ρ . Given that $\rho c^2 = \epsilon$ (from Einstein's mass-energy equivalence), acceleration for the unusual matter, $\rho + 3P/c^2 \approx -2\rho$, has a negative value. This means that there was gravitational repulsion instead of attraction, explaining the expansion.

Other properties of the Universe can be explained if the inflationary era ends by the transformation of the peculiar cold, negative-pressure matter into very hot plasma. In particular, small inhomogeneities leading to the formation of galaxies are generated by quantum fluctuations of the scalar field ψ amplified by the expansion of the Universe. The amplitude of the inhomogeneities evaluated from observations of the 3-K background radiation and from the distribution of galaxies gives information about the properties of both the scalar field ψ and

*Quantum Cosmology Workshop, 1–3 May 1987, Fermilab; IV Seminar Quantum Gravity, 25–29 May 1987, Moscow.

the present-day dark matter. This is yet another example of the mutual influence between astronomy and fundamental physics. A.D. Sakharov was the first to suggest, in 1965, although in a completely different, hydrodynamical context, the role of quantum fluctuations in the formation of the structure of the Universe.

In most previous studies of the inflationary model, two restrictive assumptions were made for the sake of simplicity: both the effective scalar field ψ driving the inflation and space-time were taken to be deterministic (classical) and homogeneous (with small inhomogeneous perturbations). Recently both these assumptions were omitted (A.A.S. and A.D. Linde, Lebedev Physical Institute). It seems that the evolution of the quasi-homogeneous large-scale part of the scalar field is essentially determined by its small-scale quantum fluctuations induced by the expansion of the Universe.

As a result, the large-scale field and space-time metric become stochastic (stochastic inflation). The evolution of the large-scale field is described by the Fokker-Planck type equation for the probability distribution $p(\psi, a)$ (a is the cosmological scale factor) with different realizations of the stochastic process in different points of space which become causally disconnected in the course of the exponential expansion during the inflationary stage. The space-time metric follows the evolution of the scalar field: it remains classical but becomes stochastic and inhomogeneous at large scales, too.

Eternal inflation

Two striking consequences follow: first, the inflationary stage as a whole appears to be eternal, having no definite end (though in some models each 'observer' spends only a finite time in it); second, the Universe after inflation is not at all homogeneous at super-large-scales which we cannot observe now (a typical value is $10^{100,000}$ cm; the present cosmological horizon is 10^{28} cm). In particular, the three-dimensional hypersurfaces of approximate homogeneity (constant-energy-density hypersurfaces) need not be the 'Cauchy' hypersurfaces of the whole space-time, and may have complicated topological structures. In other words, a picture (hypothesized earlier) of an infinite number of Friedmann universes created from one maternal de Sitter universe arises naturally. This is the price the modern inflationary theory has to pay for the explanation of approximate homogeneity on scales of the order of cosmological horizon.

This sheds new light on the problem of the origin of the Universe. It does not preclude a quantum transition creating the closed Universe with the homogeneous field ψ from nothing. But a new possibility arises: the eternal existence of

the expanding, locally homogeneous quasi de Sitter space-time supported by quantum fluctuations that has no definite three-dimensional topology and constantly creates separate Friedmann worlds.

Another fundamental problem put forward by S.W. Hawking (Cambridge University) about 10 years ago is so-called quantum unpredictability or loss of quantum coherence. It can be thought of as connected with the formation and subsequent evaporation of very small black holes during real processes, like the scattering of two energetic particles. Imagine the black hole after evaporation as a minuscule closed world geometrically detached from our space. If it carries some information inaccessible to us, the phase of the particles after scattering is not known, and the interference with simple scattering (not accompanied by black-hole creation) is incomplete. Thus, a new type of indeterminacy is superimposed on that associated with quantum gravity.

Now Hawking has developed an elegant mathematical description of this process that does not include the consideration of black holes at all. He uses the euclidean (imaginary-time) formalism in quantum gravity where the process of spontaneous detachment of a closed three-world may be described by a virtual 'wormhole' connecting two asymptotically euclidean regions, its radius being of the order of the Planck length (about 10^{-33} cm). Particles can go back and forth through this wormhole.

Is it possible to verify the suggested loss of coherence in laboratory experiments? The question is difficult because gravitational forces are very small compared with other interactions. The Planck mass 10^{-5} g, which is probably the smallest possible mass of a black hole, is much greater than the mass of a proton at rest (10^{-24} g), or even of a proton accelerated in the future supercollider (20 TeV, or 3×10^{-20} g). The proposed effect is of the order of unity for fundamental (not composite) scalar (spin-zero) particles, but we do not know any such particles. For particles of higher spin, the effect is suppressed by powers of the ratio of the particles' rest mass to the Planck mass. One prediction is that baryons (such as protons) should decay to leptons (such as electrons), the characteristic lifetime being of the order of 10^{30} yr. The problem is not only that this time is too long for the process to be detected, but that gravitational decay will be masked by much more rapid baryon decay (preserving quantum coherence) predicted by grand unified theories (with a characteristic lifetime of $10^{32} - 10^{35}$ yr).

The application of quantum unpredictability to cosmology shows that the Universe cannot be in a pure quantum state even at its beginning, so that it has to be

described not by the 'wavefunction of the Universe', but by a density matrix (Hawking; D.N. Page, Pennsylvania State University).

Fifty years ago, Dirac argued that the large difference in strength between gravitation and electromagnetism means that these two forces are of different origin. The prevalent idea now, anticipated by Einstein, is the unity of all forces of nature. But we understand better how difficult the problem is. At present, the most promising candidate for the 'theory of everything' (TOE) is superstring theory. Einstein's theory of gravity arises in its low-energy limit (with energies much less than the Planck energy 10^{19} GeV), so that general relativity is not an exact theory.

Additional curvature

The TOE should predict corrections that add more curvature (thus, they are negligible at low curvatures). The exact form of these corrections has yet to be determined. Furthermore, space-time in superstring theory is 10-dimensional, therefore six spatial dimensions must somehow compactify. Account of these properties gives rise to new multi-dimensional cosmological models (some were reported by S. Deser, Brandeis University). Another tantalizing goal is to deduce inflation from superstring theory, in particular, to derive the exact form of the potential $V(\psi)$. Up to now only a few steps have been made in this direction. One approach (D.V. Nanopoulos, CERN) is based on the so-called no-scale supersymmetry suggested by J. Ellis and collaborators a few years ago.

Last, but not least, is the problem of the vacuum energy density in flat space-time (equivalent to the problem of the cosmological constant). Vacuum is defined as the state of minimum energy density, which need not be equal to zero. The quantum approach to the problem tells us that every elementary bosonic field (such as the electromagnetic one) gives an infinite positive contribution to the energy density, and every fermionic field, an infinite negative contribution. Laboratory measurements and, with much higher precision, astronomical observations tell us that these positive and negative infinities compensate each other with fantastic precision, the energy density of the vacuum being less than 10^{-8} erg cm $^{-3}$. There is no theoretical explanation of this miraculous compensation. The best result is achieved with the above-mentioned 'no-scale' supersymmetry, which can explain why the vacuum energy density should be less than 10^{46} erg cm $^{-3}$, which corresponds to an energy scale of supersymmetry breaking of about 100 GeV. $\square$

A.A. Starobinskii is at the Landau Institute for Theoretical Physics, Academy of Sciences of the USSR, Moscow, USSR.

Information on past solar activity and geomagnetism from ^{10}Be in the Camp Century ice core

J. Beer^{*†}, U. Siegenthaler[†], G. Bonani^{*}, R. C. Finkel[‡], H. Oeschger[†], M. Suter^{*} & W. Wölfli^{*}

^{*} Institut für Mittlere Energiephysik, ETH-Hönggerberg, CH-8093 Zürich, Switzerland

[†] Physics Institute, University of Bern, CH-3012 Bern, Switzerland

[‡] Lawrence Livermore National Laboratory, Livermore, California 94550, USA

A nearly continuous record of the concentration of ^{10}Be from ice-core data is compared with the tree-ring-derived record for ^{14}C over the same period and it is found that short-term trends in these data sets provide a record of solar activity. No definitive relationship between the long-term trends and the geomagnetic dipole moment is found.

THE radioisotope ^{10}Be , with a half-life of 1.5×10^6 yr, is continuously produced in the atmosphere by the interaction of cosmic-ray particles with nitrogen and oxygen. The radioisotope production rate depends on the atmospheric cosmic-ray flux, which varies in time owing to modulation by the solar wind or the Earth's magnetic field. After a short time ^{10}Be becomes attached to aerosols and after a mean residence time of 1–2 yr is removed from the atmosphere, mainly by precipitation. Transport and deposition are strongly dependent on atmospheric circulation and mixing processes and on the distribution and the rate of precipitation. A record of the ^{10}Be concentration in precipitation therefore potentially contains information on the history of solar activity, geomagnetic field intensity and climate.

Variations of the radioisotope production rate (and therefore of the solar and terrestrial magnetic fields) have been much discussed with reference to the tree-ring record of atmospheric ^{14}C concentration^{1–3}. The geochemical behaviour of ^{10}Be is different from that of ^{14}C (^{14}C is also produced by cosmic rays in the atmosphere); in particular, ^{10}Be reflects variations in production rate much more directly than ^{14}C . On the other hand, ^{10}Be is influenced by local atmospheric processes whereas $^{14}\text{CO}_2$ is homogeneously distributed in the atmosphere, so that much may be learned from a comparison of the distribution of the two stages.

Accelerator mass spectrometry (AMS) allowed measurements to be made on relatively small samples so as to reconstruct the ^{10}Be fallout history from samples of polar ice where precipitation is preserved in a stratigraphically undisturbed way.

Several studies^{4–6,8–10} have already demonstrated the usefulness of ^{10}Be measurements in ice cores. At the Glacial/Holocene transition and during the last glaciation, strong variations of the ^{10}Be concentration in polar ice have been observed which are probably mainly due to changes in the precipitation rate^{4–6}. Measurements at the Milcent station in central Greenland have shown that during the period of solar quiescence known as the Maunder Minimum (AD 1645–1715)⁷, a period when almost no sunspots were observed, the ^{10}Be concentration was substantially higher than during other periods⁸. It appears that the 11-yr Schwabe solar cycle can also be detected in the same data⁹. First results from the Camp Century ice core, Greenland, covering the past 4,500 years have already been published¹⁰. We now describe variations from the same core during the past 10,000 years.

Experimental procedure

At Camp Century (north Greenland, 77°10' N, 82°08' W), a 1,370-m-long ice core was drilled through the ice sheet down to the bedrock¹¹, covering the last glaciation back to ~100,000 years¹². A continuous strip was cut from the core, divided into samples usually of 6 m in length and ~2 kg in mass. At two depth intervals, the Maunder Minimum (AD 1645–1715) and

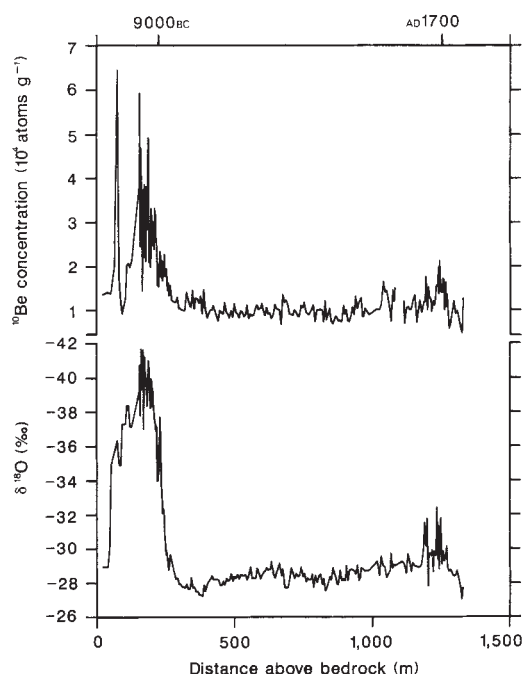


Fig. 1 ^{10}Be concentration (10^4 atoms g^{-1}) and $\delta^{18}\text{O}$ (‰) of the Camp Century ice core. The strong change of both curves at ~200 m represents the transition into the last glaciation (9000 BC). The gap in the ^{10}Be curve at ~1,100 m is caused by four missing data points. $\delta^{18}\text{O}$ is the deviation, in parts per thousand, of the $^{18}\text{O}/^{16}\text{O}$ ratio.

the Glacial/Postglacial transition (11,000 yr BP), the sampling interval was reduced to increase resolution. Samples of ^{10}Be were prepared as described previously¹³ and measured using the accelerator mass spectrometer at ETH Zürich¹⁴. Each sample was measured at least twice, resulting in a precision of 4–10%. To have the $\delta^{18}\text{O}$ values for the very same depth intervals as ^{10}Be and to compare them with the detailed record of Dansgaard and co-workers¹², $\delta^{18}\text{O}$ was measured on aliquots of each sample.

Results

Figure 1 shows the ^{10}Be concentration plotted together with $\delta^{18}\text{O}$ as a function of height above bedrock. The ^{10}Be values of the top 900 m correspond to those published earlier from Camp Century¹⁰ except for an increase by 20% due to a recalibration of the used standard³⁸. Both curves show a dramatic change at 250 m above bedrock, corresponding to the Glacial/Holocene

transition. During glacial times, the ^{10}Be concentration is considerably greater than during the Holocene and exhibits strong fluctuations which parallel the variations of $\delta^{18}\text{O}$. Similar results obtained earlier on ice cores from Antarctica^{4,6,15} and from Greenland⁵ have been attributed to changes of accumulation rate¹⁶. During the Holocene, the fluctuations of the two curves are independent except for the range of 1,200–1,250 m corresponding to the Maunder Minimum period. Here we discuss only the Holocene part of the record. First, a timescale is established, to be used to compare the ^{10}Be record with the ^{14}C data curve deduced from measurements on tree rings.

Timescale

To determine the depth-age relation of ice cores we used an ice flow model¹⁷ which uses two parameters (annual layer thickness, depth of vertical strain rate change) that are not precisely known. To allow for this, two reference depths of known ages were taken (Fig. 1). (Because the core drilling did not begin at the surface, the top of the core is not a well-defined reference point.) We have, however, been able to identify the Maunder Minimum by correlating the Camp Century ^{10}Be record with that of Milcent, which has been dated using seasonal $\delta^{18}\text{O}$ variations (see ref. 10, Fig. 2). The estimated dating accuracy is ± 20 yr. As a second reference point, we took the steep $\delta^{18}\text{O}$ increase at 1,133 m depth, already identified as the end of the Younger Dryas cold phase in Europe with a conventional ^{14}C age of $10,000 \pm 200$ yr BP¹⁹ using a correlation with oxygen isotope records from lake sediments¹⁸. For the absolute age we have to correct for variations of the atmospheric ^{14}C concentration. Extrapolating backwards the tree-ring ^{14}C series²⁰, which date back to $\sim 9,000$ yr BP, we estimate $\Delta^{14}\text{C} = 90 \pm 30\%$ at the end of the Younger Dryas, which corresponds to results obtained from Swedish varves²¹. This leads to an absolute age of $9,050 \pm 300$ BC ($11,000 \pm 300$ yr BP) for our second reference point, consistent with the result of Hammer *et al.*²² based on counting annual dust layers. The two parameters of the ice flow model were then tuned to fit the fixed points.

In Fig. 2, ^{10}Be is plotted against the resulting absolute age for the past 10,000 years (Holocene). In addition to the data from Camp Century, the results from Milcent, Greenland, covering the time AD 1180–1800 (ref. 8), have also been included after adjustment to both the depth-age scale and the mean ^{10}Be value of the Camp Century data for the same time span. Because of the thinning of the annual layers with depth, the lower (older) part of the curve is significantly expanded compared to Fig. 1.

Comparison of ^{10}Be and ^{14}C records

The Holocene ^{10}Be record in Fig. 2 shows significant variations on long timescales (for example, maxima between 6000 and 4000 BC and after AD 1000) as well as on timescales of centuries. The question is whether the observed variations are caused by changes in climate or changes of the radioisotope production rate. The ^{10}Be decrease before 7000 BC coincides with a $\delta^{18}\text{O}$ increase (see Fig. 1) and may be due to climate effects; perhaps the north Greenland ice sheet was still adapting to the transition from Glacial to Holocene conditions. After 7000 BC, there is no clear correlation between ^{10}Be and $\delta^{18}\text{O}$ (except for the range 1,200–1,250 m which corresponds to the Maunder Minimum period), suggesting that short-term ^{10}Be fluctuations are mainly caused by changes of the production rate and not by climate effects.

To test this hypothesis we have compared our results with the $\Delta^{14}\text{C}$ record from tree-ring measurements²⁰ (Fig. 3b). ($\Delta^{14}\text{C}$ is the deviation, in parts per thousand, of the ^{14}C concentration from a standard value. Around AD 1800, atmospheric $\Delta^{14}\text{C}$ was $\sim 0\%$.) The production of both ^{10}Be and ^{14}C by cosmic rays in the atmosphere depends on the energy spectrum of the primary particles. Changes of the cosmic-ray flux due to solar and geomagnetic modulation therefore cause variations of produc-

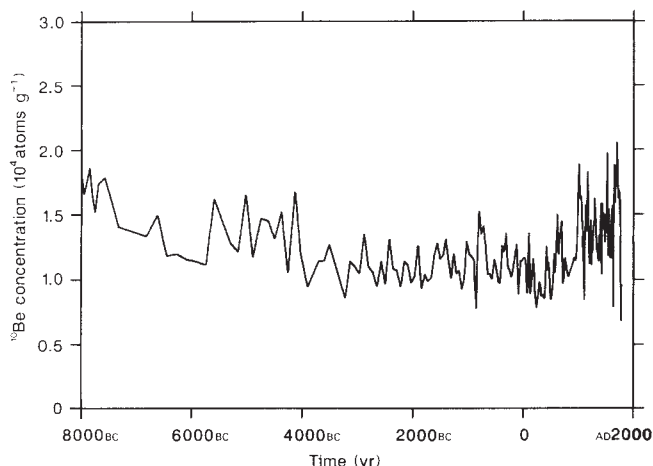


Fig. 2 ^{10}Be concentration of Fig. 1 plotted against time for the upper 1,000 m. For the period AD 1180–1800 the results from Milcent (central Greenland)⁸ are also included after corresponding adjustments of the depth-age scale and the mean value.

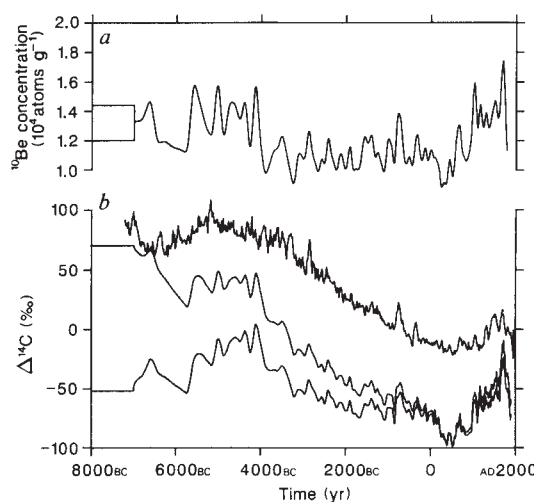
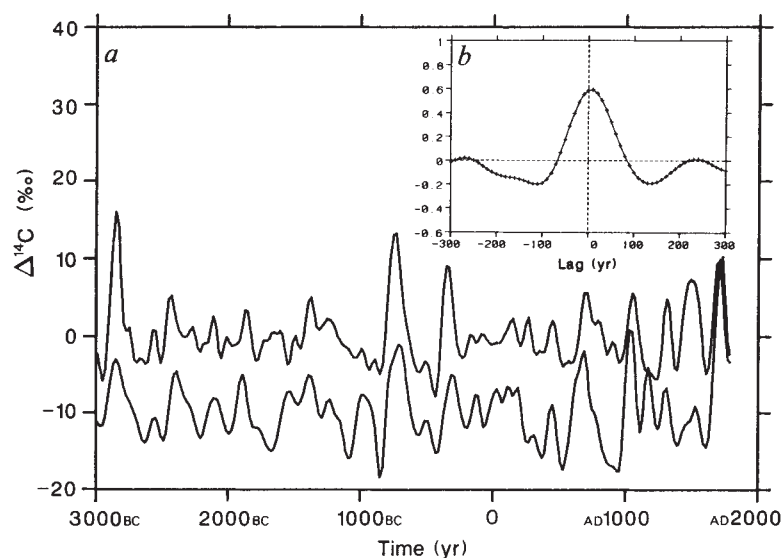


Fig. 3 *a*, ^{10}Be concentration of Fig. 2 from 7000 BC onwards, used as input function for the model calculations (units: 10^4 atoms g^{-1}). The data were slightly smoothed. The two straight lines before 7000 BC represent the two initial conditions (1.20×10^4 and 1.44×10^4 atoms g^{-1}). *b*, Measured atmospheric $\Delta^{14}\text{C}$ data²⁰ from tree rings and calculated $\Delta^{14}\text{C}$ data using the ^{10}Be data of *a* as input to the box-diffusion model. The calculations were carried out for the two different initial conditions as indicated in Fig. 3*a*. The resulting model curves were shifted by 70‰ for an easier comparison with the measured $\Delta^{14}\text{C}$ data.

tion rate of comparable relative amplitude²³. If the observed ^{10}Be fluctuations are due to changes of production rate, then similar variations should have occurred in the ^{14}C record, but if climate effects are responsible, the two records should be uncorrelated. ^{10}Be is scavenged from the atmosphere within one or two years of its creation; production rate changes are therefore reflected in the ice directly, without attenuation. By contrast, ^{14}C , present in the form of CO_2 , is first diluted by the atmospheric CO_2 mass and, with time, by increasing contributions from the other carbon pools, ocean and biosphere. The atmospheric ^{14}C concentration therefore reflects high-frequency variations of the production rate only with strong attenuation; on the other hand, it preserves a memory of previous production-rate changes. Thus, the global carbon exchange system acts as a low-pass filter for the ^{14}C production variations²⁴.

Fig. 4 *a*, Enlargement of the period 3000 BC to AD 2000 of Fig. 3*b* after removal of the long-term trends by a high-pass filter. The upper curve represents the measured $\Delta^{14}\text{C}$ variations²⁰; the lower shows calculated $\Delta^{14}\text{C}$ based on the ^{10}Be data using the higher initial conditions of Fig. 3*a*. The lower curve is shifted by 10% for easier comparison. Both curves are slightly smoothed. *b*, Cross-correlation coefficient between the measured and calculated $\Delta^{14}\text{C}$ curves (see *a*) as a function of lag time.



To compare the ice-core ^{10}Be record and the tree-ring ^{14}C data, we have simulated the low-pass filtering effect of the exchange system by means of an appropriate carbon cycle model. The model used is a modified version²⁵ of the original box-diffusion model of Oeschger *et al.*²⁶. The atmosphere is subdivided into stratosphere and troposphere, and a four-box terrestrial biosphere is included. The ocean is treated as a diffusive medium. The dynamic ocean model parameters are derived from the natural distribution of ^{14}C (ref. 27).

We assume for simplicity that the effects of atmospheric transport and mixing processes can be neglected and therefore that the ^{10}Be concentration in the Camp Century samples is directly proportional to the global production rate of ^{10}Be as well as to that of ^{14}C . Thus we use the ^{10}Be concentration record (Fig. 3*a*) as the model input function for describing the relative variations of the atmospheric ^{14}C production rate after 7000 BC. For the model results, the value of the standard ^{14}C concentration is a free parameter and is chosen such that for the period AD 0–1000 the calculated average $\Delta^{14}\text{C}$ agrees with the tree-ring mean of -14.6% . The simulation is started in 7000 BC, a time when the ^{10}Be values at Camp Century seem to be no longer influenced by the Glacial/Holocene transition. Due to the long memory of the ^{14}C system, the long-term behaviour of the calculated ^{14}C concentration is strongly dependent on the production-rate history before the start-time of the model. At the end of the last glaciation and in the early Holocene the ^{10}Be signal was strongly dominated by climate effects, and therefore the detailed isotope production rate history before 7000 BC is difficult to derive from the Camp Century record. For the calculations, we first assume that the production rate is constant before 7000 BC, with a value corresponding to the mean ^{10}Be concentration during the period 7000 BC to AD 1800 (1.2×10^4 atoms g^{-1}). The calculations were repeated for an initial production rate higher by 20%. As seen in Fig. 3*b*, this second curve fits the actual long-term $\Delta^{14}\text{C}$ trend relatively well.

Short-term variations

Short-term variations of the atmospheric ^{14}C concentration are known to have occurred fairly regularly. High-precision tree-ring measurements²⁰ have clearly confirmed the existence of fluctuations of ~ 200 -yr duration and a peak-to-peak amplitude of $\sim 20\%$ ('Suess wiggles')^{2,28–30}, as seen in Fig. 3*b*. ^{10}Be fluctuations on a similar timescale are observed, with much larger amplitude, typically $\sim 40\%$. We wish to demonstrate that these variations are caused by changes in production rate.

To test this hypothesis, we compare model-generated ^{14}C variations with the tree-ring data of the period 3000 BC to AD 1800, an interval where the time resolution of the ^{10}Be record

is good. The long-term trend was removed from both curves by applying a binomial high-pass filter. The resulting curves (Fig. 4*a*), are very similar; for example, the prominent maxima of the tree-ring record at ~ 2800 BC, 1900 BC, 700 BC, 300 BC, AD 800, AD 1100 and AD 1700 (Maunder Minimum) are also found in the ^{10}Be -based model curve.

A quantitative measure of the similarity between the two curves is obtained using the cross-correlation function (Fig. 4*b*). The correlation coefficient has a clear maximum for a shift near zero; its value of 0.58 differs from zero with very high significance ($P > 0.999$), and the amplitudes of the variations for the two curves agree well (Fig. 4*a*). The standard deviation is 3.9% for the tree-ring data and 4.4% for the model-calculated time series. All this is strong evidence that the isotope variations have a common cause, namely changes in production rate.

^{14}C variations over the past centuries are well correlated with sunspot numbers, which indicates that the short-term variations are probably due to a changing Sun³. During the Maunder Minimum, solar activity was considerably reduced, as indicated by very low numbers of observed sunspots and aurorae⁷, and consequently the solar wind was reduced. This was presumably also the case during the other periods of isotope concentration maxima.

We have limited this comparison to the time since 3000 BC, because before that the time resolution of the ^{10}Be record, which becomes coarser with increasing age, is not sufficient to resolve fluctuations of 200-yr duration. Indeed, the wiggles of the tree-ring $\Delta^{14}\text{C}$ curve before 3000 BC are not simulated well by the ^{10}Be -based model curve (Fig. 3*b*).

A by-product of the comparison is that the timescale constructed for the Camp Century ice core seems to be accurate, at least for the period from 3000 BC to the present.

Could the short-term isotope variations be related to climatic variability? Variations in snowfall rate can cause varying ^{10}Be concentration in polar ice. For Milcent station, we estimated the annual accumulation for the time span AD 1180–1800 based on the detailed $\delta^{18}\text{O}$ profile of Dansgaard *et al.* (S. Johnsen, personal communication), and found that it varies by $\sim 10\%$. The calculated ^{10}Be flux exhibits the same pattern of variation as the ^{10}Be concentration at this station; no systematic difference can be seen between flux and concentration. For the Camp Century core, a detailed history of the accumulation rate is not available, and the only parameter that can possibly give any information on climate-induced ^{10}Be variations is $\delta^{18}\text{O}$. To see whether the observed ^{10}Be variations might be connected to climatic changes, we applied a correction, based on the assumption of an inverse proportionality between concentration and accumulation rate⁶. (A change of 1% in $\delta^{18}\text{O}$ corresponds to a

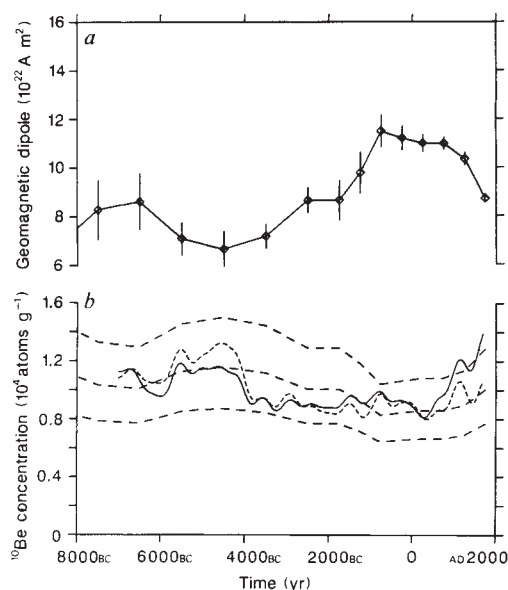


Fig. 5 *a*, Geomagnetic dipole moment M in units of 10^{22} A m² as derived from palaeomagnetic data³³. *b*, Smoothed measured ^{10}Be concentration (solid line) according to Fig. 2 and $\delta^{18}\text{O}$ -corrected ^{10}Be concentrations (dotted line). Both curves are normalized to the mean value of 1. The three dashed curves represent the expected ^{10}Be production according to the palaeomagnetic data (*a*) for three levels of solar activity (low, normal, high)²³.

20% change in ^{10}Be .) As a result, the amplitude of some of the short-term ^{10}Be fluctuations are altered, but generally, the $\delta^{18}\text{O}$ correction has a small effect. These results therefore indicate that climatic effects (at least those reflected in $\delta^{18}\text{O}$), are not the main cause of short-term ^{10}Be variations during the Holocene.

Long-term variations

The ^{10}Be record also exhibits changes on timescales of $>1,000$ yr (Fig. 2). Concentrations from ~ 6000 to 4000 BC and after AD 500 were $\sim 20\%$ higher than in the between period. These changes cannot be explained by climatic influences as expressed in $\delta^{18}\text{O}$ variations, because the $\delta^{18}\text{O}$ correction does not decrease the overall variability.

The long-term trend of the $\Delta^{14}\text{C}$ curve has often been attributed to variations of the geomagnetic dipole moment^{1,31} using the equation³² $Q/Q_0 = (M/M_0)^{-0.5}$ for the relation between isotope production rate Q and dipole moment M (Q_0 and M_0 denote present-day values). Figure 5*a* shows a compilation of a large number of palaeomagnetic data³³. These data, especially those from older periods, are subject to considerable uncertainty, as they are partly based on measurements from only a few regions and may reflect local non-dipole fields³⁴. In Fig. 5*b*, the smoothed ^{10}Be data are plotted together with the production rate calculated for the geomagnetic intensity of Fig. 5*a*, for three levels of solar activity (low, normal, high)²³. The calculated trend for constant normal solar activity shows rather poor agreement with the uncorrected ^{10}Be data (solid line). The $\delta^{18}\text{O}$ -corrected curve (dashed line) agrees better with the calculated curve.

Thus, the observed variations and those expected from palaeomagnetic data are not in good agreement, as already found earlier¹⁰. But the production rate in high latitudes is not predominantly modulated by geomagnetic field variations, so that a high-latitude record, which to some extent reflects locally produced ^{10}Be , is not ideal for studying dipole moment variations. On the other hand, a reconstruction of the geomagnetic field intensity is not without problems, so that the data in Fig. 5*a* are subject to some uncertainty³⁴.

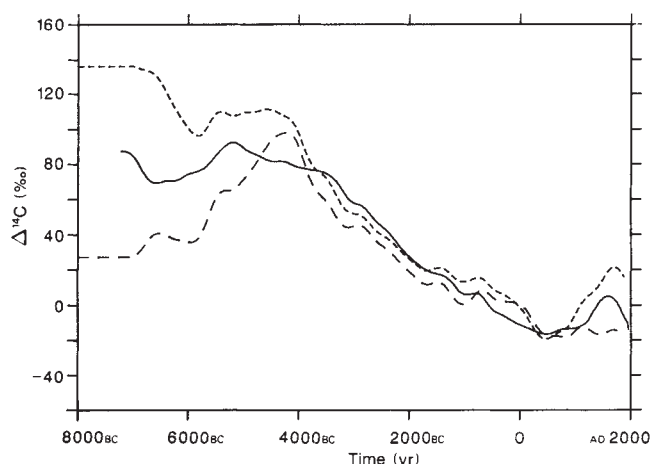


Fig. 6 Comparison of the long-term $\Delta^{14}\text{C}$ trends. Solid line, $\Delta^{14}\text{C}$ variations measured in tree rings²⁰; dotted lines, calculated $\Delta^{14}\text{C}$ variations based on the original (upper) and the $\delta^{18}\text{O}$ -corrected (lower) ^{10}Be data.

Figure 6 shows the (smoothed) model $\Delta^{14}\text{C}$ curves for the original and the $\delta^{18}\text{O}$ -corrected ^{10}Be data, both for a pre-7000 BC ^{10}Be concentration of 1.44×10^4 atoms g⁻¹. Although the decrease from 4000 BC to AD 500 is similar, there are clear differences before 4000 BC and after AD 500. The curve obtained from the $\delta^{18}\text{O}$ -corrected ^{10}Be data does not simulate the tree-ring record as well for these periods, and therefore does not support making the $\delta^{18}\text{O}$ correction on the ^{10}Be record.

When discussing the long-term trend it is important to be aware that the ^{14}C system has a long memory; the model-simulated trends (Figs 3, 6) depend strongly on the initial conditions. For a pre-7000 BC production rate higher by 20% (upper curve in Fig. 3*b*), the long-term tree-ring trend is relatively well simulated. This means that the decreasing trend does not necessarily imply a slowly decreasing production rate after 4000 BC, but could well be the consequence of $\sim 20\%$ higher production rate before 7000 BC. Because the global ^{14}C inventory adjusts to long-term production rate changes with a half-life of 5,730 yr, a new quasi-steady state would not be reached yet. Therefore the mean pre-7000 BC ^{14}C concentration would not have been $+200\%$, but $\sim +140\%$. (Note that 0% corresponds to the atmospheric value at $\sim$ AD 1800.)

It has been thought that the long-term ^{14}C trend might be caused by a gradual increase in the speed of ocean circulation³⁵, which leads to a faster vertical ocean circulation and therefore to an increased release of 'old', ^{14}C -depleted CO_2 into the atmosphere²⁴. Simultaneous ^{14}C measurements on surface-dwelling and bottom-dwelling foraminifera³⁶ on a deep-sea sediment core do not support the suggested speed-up of the circulation. It seems, however, well established that during glacial time the rate of deep-water formation in the North Atlantic was reduced, which must have caused higher atmospheric ^{14}C concentrations by $\sim 50\text{--}100\%$. A model simulation²⁵ indicates that the atmospheric ^{14}C concentration adapts within $<1,000$ yr to a change towards a faster ocean circulation such as that which occurred $\sim 13,000$ years ago. This, and the foraminiferal ^{14}C data, show that the long-term $\Delta^{14}\text{C}$ trend cannot be an after-effect of an internal change in the carbon system at the end of glacial time.

Although no final conclusions can be drawn regarding the origin of the $\Delta^{14}\text{C}$ long-term trend, the available information is sufficient to advance a hypothesis different from the often discussed picture^{1,31} that this trend is the direct result of changing geomagnetic dipole moment. The comparison of ^{10}Be and ^{14}C suggests that the isotope production rate was higher by 20% during the last 10,000–15,000 yr of the ice age, leading to correspondingly higher ^{14}C concentrations in all carbon reservoirs

($\Delta^{14}\text{C} = 140\%$ in the atmosphere). If the production rate changed to the present-day value some time in the early Holocene, the decreasing trend between 5000 and 0 BC would reflect the exponential decay of ^{14}C with a half-life of 5,730 yr. The maximum of the ^{14}C concentration between 6000 and 4000 BC and the increase after AD 0 could have been caused by a changing geomagnetic dipole moment, as indicated by the palaeomagnetic data³³.

The late glacial production rate is therefore essential for the interpretation of the long-term trend in ^{14}C concentration. Unfortunately, the ^{10}Be signal at that time is completely masked by climate effects. There are, however, ^{14}C dates from varves²¹ indicating that the atmospheric ^{14}C level was higher at the end of the last glaciation. These results are in agreement with data indicating that the geomagnetic field was weaker during the period 50,000–10,000 yr BP³⁷.

Conclusions

The comparison of the Camp Century ^{10}Be record with $\delta^{18}\text{O}$ and $\Delta^{14}\text{C}$ shows that ^{10}Be measurements on polar ice cores give valuable information. Strong climatic changes such as the transi-

tion from glacial to interglacial times or climatic fluctuations during glaciation are clearly visible in the ^{10}Be record, but climatic variations do not seem to have had much influence during the Holocene. The good correlation between the main short-term variations of both the ^{10}Be record in ice and the ^{14}C record in tree-rings over the past 5,000 yr strongly supports the explanation that these fluctuations are caused by solar modulation of the Galactic cosmic-ray flux. The comparison of ^{10}Be and ^{14}C also makes possible the refinement of ice-core dating using the technique of wiggle matching.

Less clear conclusions can be drawn regarding the long-term isotope variations. The ^{10}Be data do not give support to the hypothesis that the observed slow ^{14}C decrease was due to a gradual geomagnetic field change. They suggest, rather, that the ^{14}C trend might be the result of a 20% higher production rate during the last 10,000–15,000 years of the ice age.

We thank C. C. Langway, B. Kapuza and M. Andr  e for help with the ice sampling, K. H  nni for measuring $\delta^{18}\text{O}$, H. J. Hofmann, E. Morenzoni and M. Nessi for help during the ^{10}Be measurements and C. U. Hammer, H. Clausen and B. Stauffer for useful discussions. This work was supported financially by the Swiss NSF.

Received 20 July; accepted 7 December 1987.

1. Bucha, V. in *Radiocarbon Variations and Absolute Chronology*, 501–513 (Almquist & Wiksell, Stockholm, 1970).
2. Suess, H. in *Radiocarbon Variations and Absolute Chronology*, 595–605 (Almquist & Wiksell, Stockholm, 1970).
3. Stuiver, M. & Quay, P. D. *Science* **207**, 11–19 (1980).
4. Raisbeck, G. M. et al. *Nature* **292**, 825–826 (1981).
5. Beer, J. et al. *Ann. Glaciol.* **5**, 16–18 (1984).
6. Yiou, F., Raisbeck, G. M., Bourles, D., Lorius, C. & Barkov, N. I. *Nature* **316**, 616–617 (1985).
7. Eddy, J. A. *Science* **192**, 1189–1201 (1976).
8. Beer, J. et al. in *Proc. 18th Int. Cosmic Ray Conf. Bangalore* vol. 9 (eds Durgaprasad, N., Ramadurai, S., Ramana Murthy, P. V., Rao, M. V. S. & Sivaprasad, K.) 317–320 (P. V. Ramana Murthy, Bombay, 1983).
9. Beer, J. et al. *Nucl. Instrum. Meth.* **B10/11**, 415–418 (1985).
10. Beer, J. et al. *Nucl. Instrum. Meth.* **B5**, 380–384 (1984).
11. Ueda, H. T. & Garfield, D. E. *U.S. Army Cold Regions Res. Engng Lab. spec. Rep.* 126 (1968).
12. Johnsen, S. J., Dansgaard, W., Clausen, H. B. & Langway, C. C. *Nature* **235**, 429–434 (1972).
13. Beer, J. et al. *Radiocarbon* **25**, 269–278 (1983).
14. Suter, M. et al. *Nucl. Instrum. Meth.* **B5**, 117–122 (1984).
15. Raisbeck, G. M. et al. *Nature* **326**, 273–277 (1987).
16. Raisbeck, G. M. & Yiou, F. *Ann. Glaciol.* **7**, 138–140 (1985).
17. Hammer, C. U. et al. *J. Glaciol.* **20**, 3–25 (1978).

18. Oeschger, H. et al. *Geophys. Monogr.* **29**, 299–306 (1984).
19. Mangerud, J., Andersen, S. T., Berglund, B. E. & Donner, J. J. *Boreas* **3**, 109–128 (1974).
20. Stuiver, M. & Kra, R. S. *Radiocarbon* **28**, 805–1016 (1986).
21. Stuiver, M., Kromer, B., Becker, B. & Ferguson, C. W. *Radiocarbon* **28**, 969–979 (1986).
22. Hammer, C. U., Clausen, H. B. & Tauber, H. *Radiocarbon* **28**, 284–291 (1986).
23. Blinov, A. *Proc. NATO adv. Res. Workshop Durham*, 1987 (eds Stephenson, R., Wolfendale, A. & Eddy, J. A.) (Reidel, Dordrecht, in the press).
24. Siegenthaler, U., Heimann, M. & Oeschger, H. *Radiocarbon* **22**, 177–191 (1980).
25. Siegenthaler, U. & Beer, J. *Proc. NATO adv. Res. Workshop Durham* 1987 (eds Stephenson, R., Wolfendale, A. & Eddy, J. A.) (Reidel, Dordrecht, in the press).
26. Oeschger, H., Siegenthaler, U., Gugelmann, A. & Schotterer, U. *Tellus* **27**, 168–192 (1975).
27. Siegenthaler, U. *J. geophys. Res.* **88**, 3599–3608 (1983).
28. Damon, P. E., Lermann, J. C. & Long, A. A. *Rev. Earth planet. Sci.* **6**, 6457–6494 (1978).
29. Neftel, A., Oeschger, H. & Suess, H. *Earth planet. Sci. Lett.* **56**, 127–147 (1981).
30. Sonett, C. P. *Rev. Geophys. Space Phys.* **22**, 239–254 (1984).
31. Damon, P. E. & Linick, T. W. *Radiocarbon* **28**, 266–278 (1986).
32. Elsasser, W., Ney, E. P. & Winckler, J. R. *Nature* **178**, 1226–1227 (1956).
33. McElhinny, M. W. & Senanayake, W. E. *J. Geomagn. Geoelectr.* **34**, 39–51 (1982).
34. Barton, C. E., Merrill, R. T. & Barbetti, M. *Phys. Earth planet. Inter.* **20**, 96–110 (1979).
35. Lal, D. *Geophys. Monogr.* **32**, 221–233 (1985).
36. Andr  e, M. et al. *Clim. Dyn.* **1**, 53–62 (1986).
37. Barbetti, M. & Flude, K. *Nature* **279**, 202–205 (1979).
38. Hofmann, H. J. et al. *Nucl. Instrum. Meth.* **B29**, 32–36 (1987).

A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons

David Zipser* & Richard A. Andersen†

* Institute for Cognitive Science, University of California, San Diego, La Jolla, California 92093, USA

† Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

Neurons in area 7a of the posterior parietal cortex of monkeys respond to both the retinal location of a visual stimulus and the position of the eyes and by combining these signals represent the spatial location of external objects. A neural network model, programmed using back-propagation learning, can decode this spatial information from area 7a neurons and accounts for their observed response properties.

THIS article addresses the question of how the brain carries out computations such as coordinate transformations which translate sensory inputs to motor outputs. Visual inputs are collected in the coordinate frame of the retina on which the visual environment is imaged, but motor movements such as reaching are made to locations in external space. Changes in eye position will alter the retinal locations of targets while their spatial locations remain constant. As a result, visual inputs must be transformed from retinal coordinates to coordinates that specify

the location of visual objects with respect to the body to perform accurately directed movements.

Lesions to the posterior parietal cortex in monkeys and humans produce profound spatial deficits in both motor behaviour and perception^{1–5}. Humans with lesions to this area can still see but they appear to be unable to integrate the position of their bodies with respect to visual inputs. The lesion data further suggest that it is the inferior parietal lobule, which comprises the posterior half of the posterior parietal cortex,

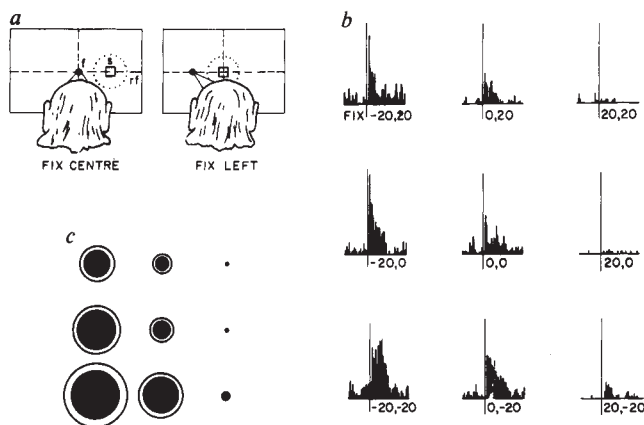


Fig. 1 *a*, Experimental protocol for determining spatial gain fields, with the projection screen viewed from behind the monkey's head. These experiments were carried out several years before the start of the modelling described here¹⁰, but some of the data are being presented for the first time. To determine the effect of eye position, the monkey with head fixed, fixates on a point, *f*, at one of 9 symmetrically placed locations on the projection screen. The stimulus, *S*, is always presented at the same retinal location, chosen as the maximum-response zone of the retinal-receptive field. The stimulus consists of 1- or 6-degree diameter spots flashed for 500 ms. Each measurement is repeated 8 times. *b*, Peri-stimulus histograms of a typical gain field determination. The nine histograms are located in the same relative positions as the fixations that produced them. The vertical line indicates the time of visual stimulus onset. *c*, A graphic method for illustrating these data in which the diameter of the darkened inner circle, representing the visually evoked gain fields is calculated by subtracting the background activity recorded 500 ms before the stimulus onset from the total activity during the stimulus. The outer circle diameter, representing the total response gain fields, corresponds to the total activity during the stimulus. The annulus diameter corresponds to the background activity that is due to an eye-position signal alone, recorded during the 500 ms before the stimulus presentation.

which is involved in spatial processes.

Anatomical and physiological experiments in macaque monkeys indicate that the inferior parietal lobule contains at least four separate cortical fields. Area 7a contains visual and eye-position neurons⁶⁻¹⁰; area 7b contains somatosensory and reach-related cells^{9,11}; area MST contains visual motion and smooth pursuit eye movement activity (refs 12-14; R. H. H. Wurtz and W. T. Newsome, personal communication); and area LIP contains visual and saccade-related activity^{15,16}. It has been proposed^{10,17,18} that the area most likely to perform spatial transformations is area 7a. Most of the cells in this region were found to receive a convergence of both eye-position and retinal signals. The interaction between eye-position and visual responses was non-linear, and in most cases the visual response could be modelled as a gain that was a function of eye position multiplied by the response profile of the retinal receptive field. Thus the visual receptive field remained retinotopic, but the magnitude of the response was modulated by eye position. This modulation can be shown to produce a tuning for the location of targets in head-centred coordinates that is eye-position-dependent; the cells will fire most for a particular location in craniotopic space, but only when the eyes are at the appropriate positions in the orbits. No cells were found that coded target location over all eye positions (eye-position-independent coding), indicating that this information can only be contained in the pattern of activity of a population of neurons. Here we describe a neural network model that shows how eye-position-independent location can be extracted from a population of area 7a neurons. The model also reproduces the non-linear interactions of eye position and retinal position information seen in actual area 7a neurons, and

demonstrates response properties, such as large receptive fields, which are strikingly similar to those observed in single-unit recording studies.

The neural modelling technique we use differs significantly from most previous approaches, which first found an algorithm to accomplish the computation, and then specified neural models to implement the algorithm. Our approach is based on the use of a neural network training procedure, called 'back propagation', which can programme artificial neural networks to compute arbitrary functions (refs 19-23; S. R. Lehky and T. J. Sejnowski, personal communication). Unlike computer programming, programming by training uses only examples of input and output. This means that it is not necessary to know in advance the algorithm that the network will use; the learning process will discover an appropriate algorithm. Back propagation networks have internal or hidden units that are free to take on the response properties to best accomplish the computational task being learned. It is the properties of these hidden units that we find resemble those of cortical neurons. The back-propagation procedure accomplishes learning by adjusting the strengths of the synapses within the network.

Experimental results from area 7a

The experimental data that must be accounted for by any model of area 7a were collected previously in an extensive series of studies with awake, unanaesthetized monkeys¹⁰. Here we describe new analyses of these data that facilitate a comparison of area 7a cells with the units generated by training the network model.

Three of the major classes of area 7a neurons are of interest here: the eye-position cells responding to eye-position only (15% of all cells sampled from area 7a), the visual cells responding to visual stimulation only (21%), and the spatially tuned cells responding to both eye position and visual stimulation (57%). Neurons in the first two classes presumably represent the eye position and retinal location information available to area 7a as input. The interaction of eye position and visual information found in the third class of cells produces a representation for the head-centred location of visual targets that is eye-position-dependent.

The experimental protocol involved recording neuronal activity extra-cellularly from awake, unanaesthetized monkeys trained in various visuospatial tasks¹⁰ (see Fig. 1). The eye-position sensitivity was tested by having the animal fixate on a small point at different eye positions, with the head fixed in otherwise total darkness. The eye-position cells typically showed a linear increase in activity for a range of horizontal or vertical eye positions, although some cells showed more complex eye-position coding. An ensemble of 30 eye-position unit responses is shown in Fig. 4c. The receptive fields of the visual cells were tested by flashing a spot stimulus at different locations in the visual field while the animal fixated on a target at a single eye position. Surfaces were fit to these data points using a gaussian interpolation. These cells typically had large receptive fields equally distributed across the visual field for the population of neurons with a single peak of activity. The shape of the receptive fields approximated a symmetrical gaussian with a $1/e$ width of 15 degrees (see legend to Fig. 4).

The spatially tuned neurons were the largest group and showed a convergence of eye position and retinal position information. The receptive fields are very large (often over 80 degrees in diameter) and have one or more peaks that form a smoothly changing, hilly landscape. A set of 12 retinal receptive fields from spatially tuned cells, arranged according to peak eccentricity and complexity, is shown in Fig. 2.

As mentioned above, the evoked visual response of spatially tuned neurons varies as a function of eye position. This effect was examined by collecting data under the condition in which the visual stimulus always appears at the peak location in the retinal receptive field, but with the animal fixating at nine

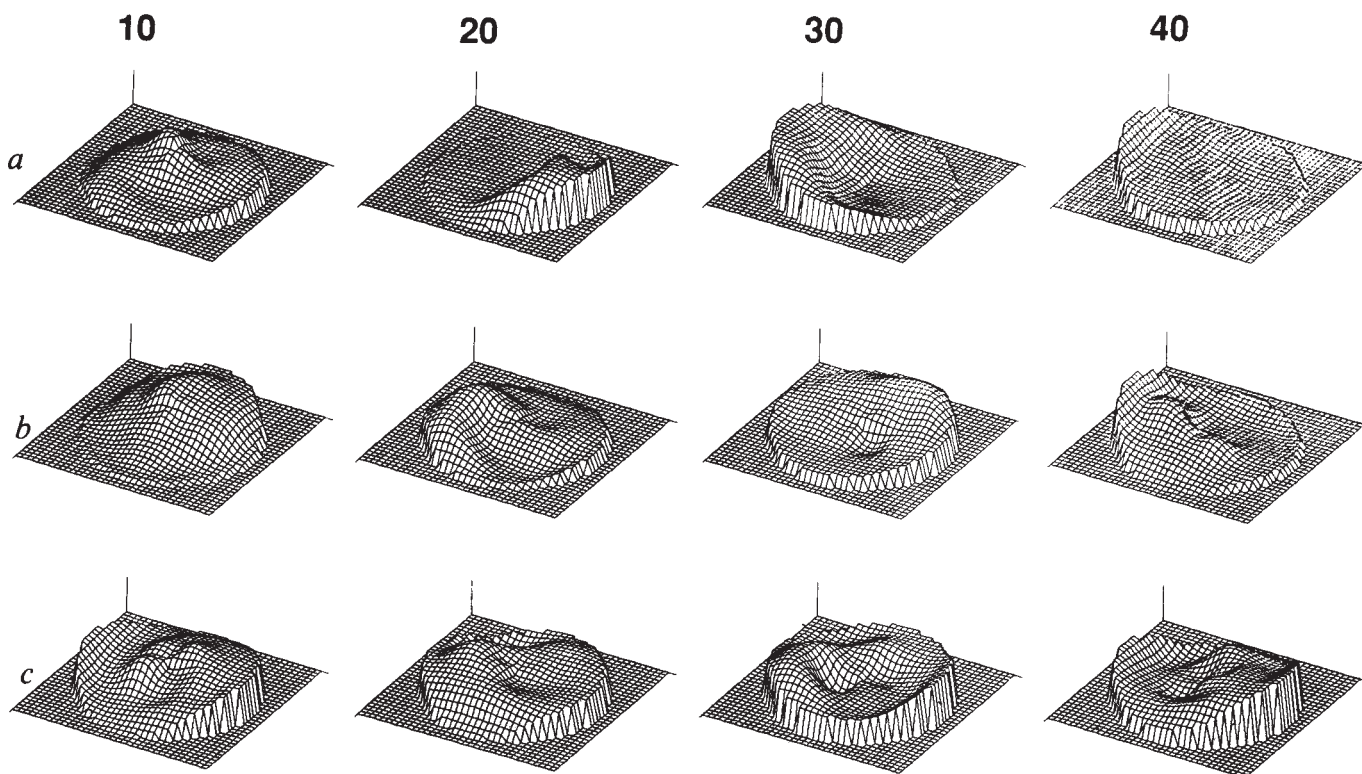


Fig. 2 The receptive fields of spatially tuned neurons from area 7a, arranged in rows with the eccentricity of the field maxima increasing to the right, and in columns with the complexity of the fields increasing downwards. Receptive fields were sampled at 17 radially spaced points, with one sample taken at the centre of the field, and four samples taken on each of four circles of radius 10, 20, 30 and 40 degrees. All the fields in row *a* have single peaks. Those in row *b* have a single large peak but some complexities in the field. The fields in row *c* are the most complex with multiple peaks. The data have been normalized so that the highest peak in each field is the same height.

different eye positions¹⁰. These plots are referred to as spatial gain fields. Figure 1 demonstrates the experimental protocol for mapping spatial gain fields.

The majority of the spatial gain fields are roughly planar. This planar behaviour is evident in Fig. 1c, where the darkened inner circles are proportional to the magnitude of the visual evoked response, the outer circle diameter to the total activity during the flashed stimulus, and the annulus diameter the background activity due to the eye position alone. The data for the visually evoked response in Fig. 1c can be fitted by a plane tilted up for eye positions to the left and also tilted up for downward eye positions. Analysis using linear regressions in the two dimensions of eye position indicated that the gain fields for the visually evoked activity (represented by the dark inner circles) were planar, or had a large planar component, for 55% of the neurons. Interestingly, 80% of the total response gain fields (represented by the outer circles) were planar or largely planar.

A useful way to further characterize the nonlinear combinations of eye and retinal information is to compare the contributions of each to the total response of the cells. This can be done simply by comparing the dark inner circles, which represent the visual contribution to the response, with the white annuli, which represent the eye position contribution to the response. This comparison shows three basic types of gain fields. For 28% the background and evoked activities change in a parallel fashion (Fig. 3b, e, f). In most of the cells (43%), the evoked activity changes with eye position while the background activity, if any, remained constant (Fig. 3a, c, d); three-quarters of these cells had very low or undetectable background activity. The remaining 28% of the neurons showed the interesting property that the background and evoked activities changed in different directions, so that the activity of either alone was grossly non-planar, but the overall activity was planar (Fig. 3g, h, i).

The neural network model

We used a three-layer network (illustrated in Fig. 4) that was trained to map visual targets to head-centred coordinates, given any arbitrary pair of eye and retinal positions. The first, or input, layer has two sections, an array of units on which the visual stimulus is represented, and a set of units representing eye position. The second layer consists of the hidden units, which map the input to the output. Each hidden unit receives input

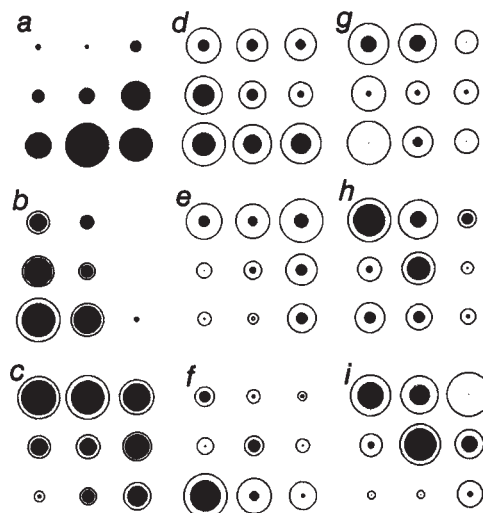
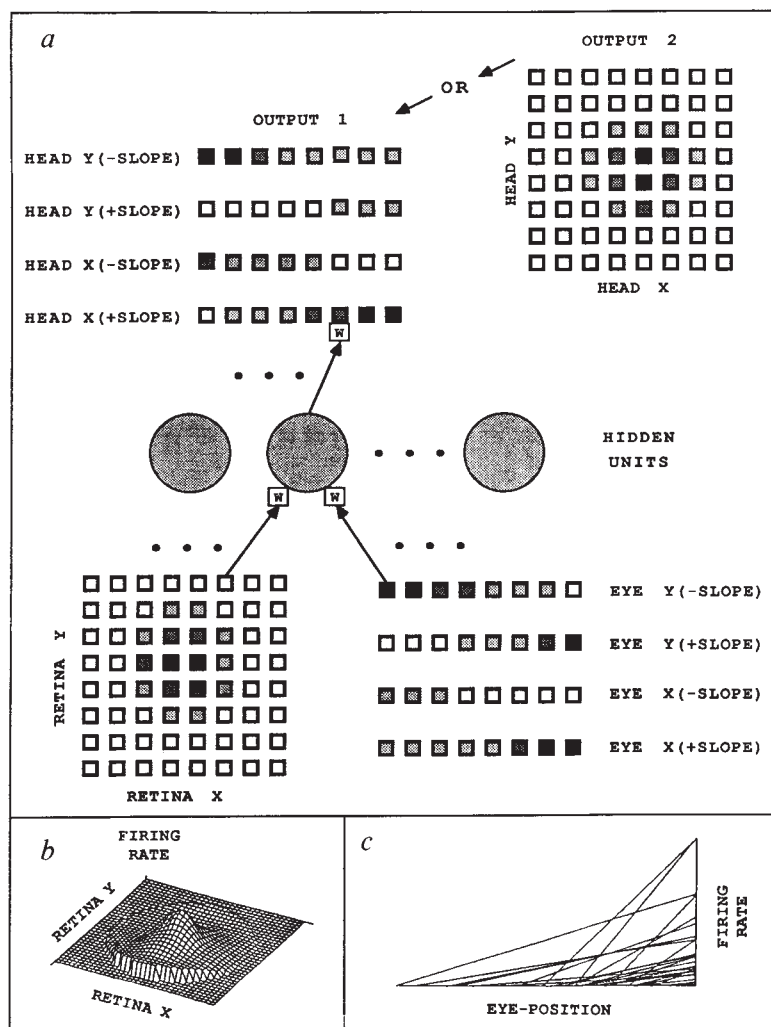


Fig. 3 The spatial gain fields of 9 neurons (a-i) from area 7a in the format of Fig. 1c.

Fig. 4 *a*, Back-propagation network used to model area 7a. The visual input consists of 64 units with gaussian receptive fields with $1/e$ widths of 15 degrees. The centre of each receptive field occupies a position in an 8 by 8 array with 10 degree spacings. The shading represents the level of activity for a single-spot stimulus, with darker shading representing higher rates of activity. The units have been arrayed topographically for illustrative purposes only; this pattern is not an aspect of the model as each hidden unit receives input from every one of the 64 retinal input units. The eye position input consists of 4 sets of 8 units each with two sets coding horizontal position (one for negative slope and one for positive slope) and two sets coding vertical position. Shading represents the level of activity. The intercepts have been ordered for illustrative purposes only and do not represent information available to the hidden layer. Each eye position cell projects to every unit in the hidden layer. Two output representations were used; the gaussian output format is shown on the right and the monotonic format on the left. The gaussian format units have gaussian shaded receptive fields plotted in head-centred coordinates. They have $1/e$ widths of 15 degrees and are centred on an 8 by 8 array in head coordinate space with 10 degree spacings. The monotonic format units have firing rates that are a linear function of position of the stimulus in head-centred coordinates. There are four sets of 8 units with two sets of opposite slope for vertical position and two sets for horizontal position in head-centred coordinates. Again, shading represents the degree of activity and the topographic ordering is for illustrative purposes only. The small boxes containing *W* indicate the location of the synapses whose weights are trained by back propagation. Each hidden unit projects to every cell in the output layer. The output activity of the hidden and output layer units is calculated by the logistic function: $\text{output} = 1/(1 + e^{-\text{net}})$, where $\text{net} = (\text{weighted sum of inputs}) + \text{bias}$. The arrow for the connections represents the direction of activity propagation; error was propagated back in the opposite direction. The back-propagation procedure guarantees that the synaptic weight changes will always move the network towards lower error by implementing a gradient descent in error in the multi-dimensional synaptic weight space.

b, Area 7a visual neuron receptive field with a single peak near the fovea. Receptive fields were plotted using the same method as in Fig. 3. Visual cells that had no eye-position-related activity or modulation of their responses by eye position were used to model the retinal input to the network. *c*, A composite of 30 area 7a-eye-position units, whose firing rates are plotted as a function of horizontal or vertical eye deviation. The slopes and intercepts are experimental values for eye-position neurons.



from every input unit and projects to every unit in the third output layer. The output of each unit of the hidden and output layers is computed as an S-shaped (logistic) function of the synaptic strength weighted sum of its inputs, plus a bias term. The training paradigm uses back propagation learning, which consists of choosing an input and desired output, applying the input to the first layer of the network and propagating the activity it generates through the network to the output units. The actual output is then subtracted from the desired output to generate an error. This error is used to adjust the weights of synapses on the output layer units and hidden layer units in a manner prescribed by the back propagation procedure¹⁹. Training begins with all weights randomized, resulting in large errors, and the training cycle is repeated until the error is reduced to desired levels.

The retinal position and eye position inputs to the network are modelled using characteristics of the cells in the posterior parietal cortex that respond to visual stimuli only and eye position only. The visual input consisted of 64 gaussian-shaped receptive fields, with $1/e$ widths of 15 degrees and with each peak separated by 10 degrees in an 8 by 8 array. The eye position input consisted of four sets of 8 units with single sets for positive and negative slopes for horizontal and vertical eye position.

We used two representations of location in head-centred coordinates at the output layer. One (output 2 in Fig. 4a) was

a gaussian format in which each unit had a gaussian receptive field similar to the representation of the retinal input, but coding location in head-centred rather than retinal coordinates. The other (output 1 in Fig. 4a) was a monotonic format, in which the activity of each neuron is a linear function of the location of the stimulus in head-centred coordinates. For the gaussian format, a 64-unit array similar to the retinal input array was used and for the monotonic format, a 32-unit array similar to the eye position input array was used. The gaussian and monotonic formats were chosen because they represent the most common types of coding formats found for brain cells. Also the monotonic format has the interesting feature that it has the same representation as the eye-position code at the input. Thus, if the animal foveates the visual stimulus, the resulting eye-position signal could be used as the teacher to indicate the correct location of the stimulus in head-centred coordinates.

The model network was trained using randomly selected pairs of input eye positions and retinal positions. The teacher signal (desired output) used to train the output units was the true spatial location in head-centred coordinates implied by the inputs, and was represented in either the monotonic or gaussian format. The network trained quickly: after ~1,000 trials, accuracies equivalent to the distance between retinal unit centres were reached. When training was continued, error continued to decrease, but at a lower rate.

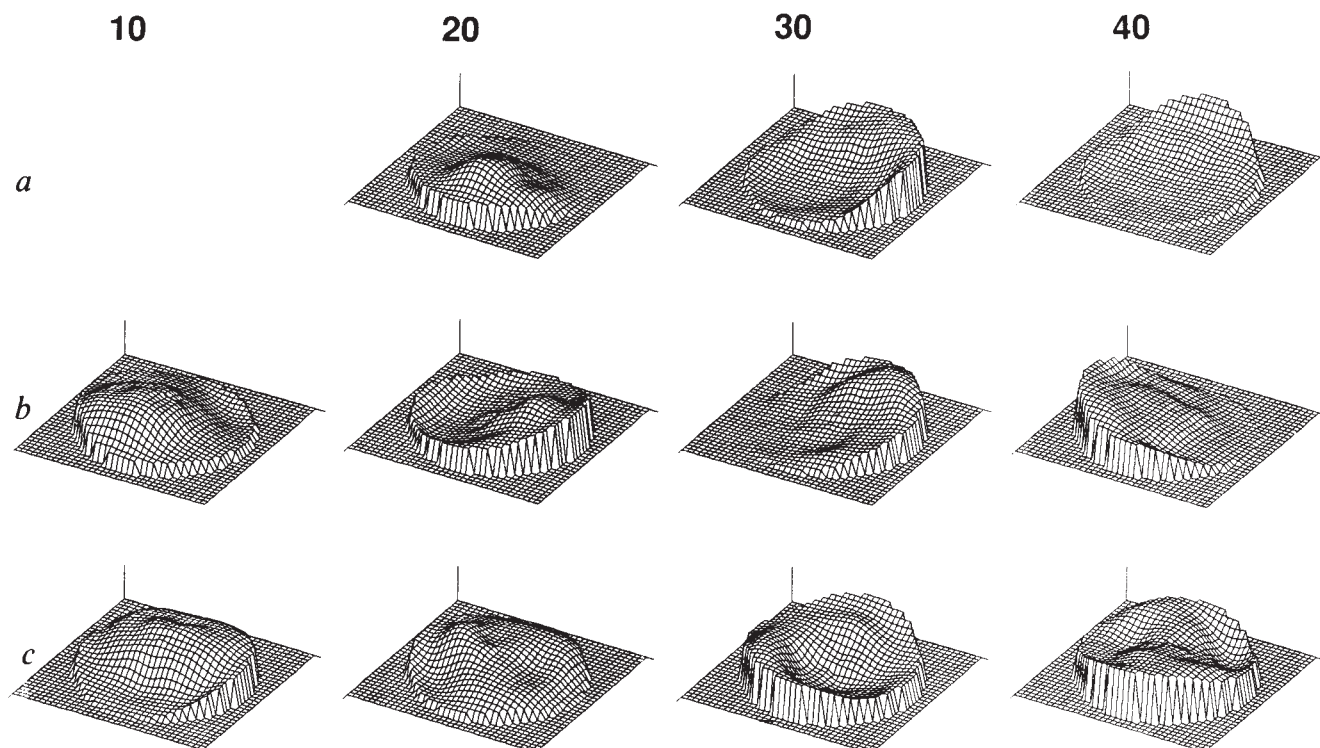


Fig. 5 Hidden unit retinal receptive fields generated by the back propagation model. These plots were generated by holding the eye-position input to the network constant and simulating visual stimulation at the same 17 retinal positions used in the experiments on area 7a. The hidden unit activities were normalized and plotted in the same way as the experimental data shown in Fig. 3. The data shown here are from a series of 4 training sessions using networks with 25 hidden units and the monotonic format output. Similar results were obtained for the gaussian format output. All the fields, except for C-10, C-20 and C-30, are from networks that have received 1,000 learning trials. The remaining three are from untrained networks, resulting only from the random synaptic weights assigned at the start of a training run. Very complex fields are only rarely found in trained networks. No hidden unit with a single peak at 10 degrees appeared in this data set and such units are very rare in trained networks. No spatially tuned neurons with central receptive fields were found in area 7a, and no such fields appeared in the trained model. But central receptive fields are found among the visual neurons in 7a, and this kind of unit was among those used as input to the model network.

Agreement of model with experiment

To evaluate the model, we first compare the experimental and model retinal receptive fields. The model receptive fields were categorized according to their complexity, and the eccentricity of their activity maxima (Fig. 5), as had been done previously

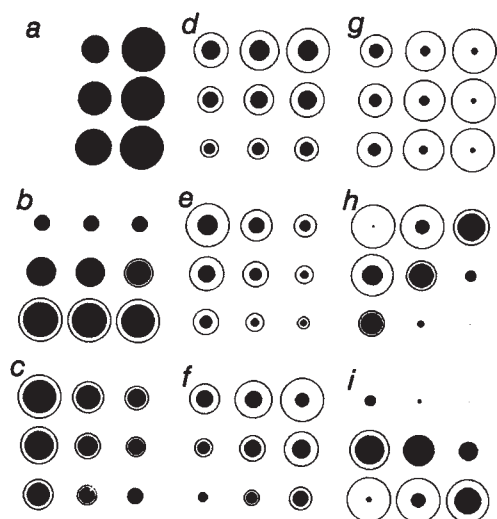


Fig. 6 Hidden unit spatial gain fields generated by the model network. Fields a-f were generated using the monotonic format output; the rest used the gaussian format output.

for the experimental receptive fields (Fig. 2). Comparison of the top lines in Figs 2 and 5 shows that the trained models generate single-peak receptive fields resembling those observed experimentally at all eccentricities except 10 degrees. The fully trained model also produces moderately complex fields like those found in line 2, but rarely produces receptive fields as complex as those in the bottom line of Fig. 2. This kind of highly complex field is not distinguishable from the untrained model receptive fields shown in the bottom of Fig. 5. The comparison process contains an element of subjectivity, but it demonstrates that the trained model generates retinal receptive fields remarkably similar to the experimentally observed fields. The gain fields generated by the model are shown in Fig. 6. All the total-response gain fields, whether generated by the monotonic or the gaussian output format, were planar in shape. This result compares with 80% of the experimental fields in this planar class. When the model gain fields are examined in more detail, taking into account the non-linearities of the visual response fields, there are significant differences between the eye position and retinal output formats. For example, when trained with the monotonic output format, 67% of the visual response gain fields were planar, but when trained with the gaussian output format only 13% fall in this class. These figures compare with 55% in this class for the experimental data. The irregular visual response gain fields generated by the gaussian output format are more radically irregular than those generated by the monotonic output format. Thus it appears that to account for the details of the visual response gain fields, it may be necessary to use both types of output representation. It should be pointed out, however, that whereas the visual receptive fields and the total gain fields

were virtually unaffected by changing parameters of the model, the visual response gain fields were very sensitive to parameters such as threshold value and output representation. The number of hidden units had little effect, giving similar results for simulations ranging from 9 to 36 hidden units.

The striking similarity between model and experimental data certainly supports the conjecture that the cortex and the network generated by back propagation compute in similar ways. As back propagation generates optimal solutions which produce the least error, these results also suggest that the brain chooses optimal solutions with respect to error. The similarity of the model and test results raises the question of what physiological mechanisms could subserve this equivalence. Presently the back-propagation paradigm is structured at a level higher than implementation, and obviously cannot be applied literally to the brain, because information does not travel backwards rapidly through axons. That the back-propagation method appears to discover the same algorithm that is used by the brain in no way implies that back propagation is actually used by the brain. One approach to understanding the physiological significance of these results is to generate models incorporating features found in the brain, such as Hebbian-like learning at synapses and reciprocal back-projection pathways for propagating error, and determine whether these models generate similar results. In this regard it is interesting to note that all cortico-cortical and thalamocortical connections have reciprocal feedback pathways.

An important consideration in interpreting the results is how closely does the model response actually resemble the cortical data? This is a complex issue because there is error in the experimental data, and additional errors introduced by the interpolation process used to produce the full-field views of the receptive fields. Examination of the magnitudes of these errors indicates that they could not account for the various receptive field types, or eye-position gain fields observed experimentally and in the model. The methods of comparison between model and experiment that we have used are to some degree subjective. Perhaps a more objective comparison procedure will eventually be developed, but it is unlikely that, given the complex nature of the data to be compared, any such technique will

substantially alter our conclusions concerning the degree of similarity between model and experimental data.

From the physiological perspective these experiments raise the possibility that the posterior parietal cortex learns to associate body position with visual position to localize accurately the position of objects with respect to the body. This idea about the importance of learning is reasonable considering that it would not be practical for spatial representations to be hard-wired because the body dimensions change during development. Furthermore, adaptation experiments show that distortion of space with prisms leads to rapid recalibration, suggesting that these representations are still plastic in adults. As the model, by definition, does not have a topographic organization to localize in space, it shows that the brain does not need a topographic organization to localize in space. The organization of the network is not a product of the spatial position of the cell bodies, but rather is contained in the pattern of the weights of the synaptic connections.

Finally, there is the question of where the output units of the model could exist in the brain. One possible location would be areas that receive projection from area 7a. Although eye position effects on visual responses have been described at several locations in the brain (refs 24–26; S. Funahashi, C. J. Bruce, P. S. Goldman-Rakic; and R. Lal, M. J. Friedlander, personal communication), an eye-position-independent coding has yet to be unequivocally demonstrated. But it is also possible that the final spatial output could only exist in the behaviour of the animal. For example, the muscles innervating the eye or limb are broadly tuned, and the position of the eye or limb is coded in a distributed fashion over the activity of several muscles. Thus the final spatial output may not exist in any single cell in the brain, but rather might be found only in the pointing of the eye or finger accurately to a location in space.

We thank David Rumelhart, Francis Crick and Emilio Bizzi for stimulating discussion during the development of this model and Carol Andersen for editorial assistance. D.Z. was supported by grants from the System Development Foundation, the AFOSR and the Office of Naval Research. R.A.A. was supported by the NIH, the Sloan Foundation and Whitaker Health Sciences Foundation.

Received 19 October 1987; accepted 19 January 1988.

1. Critchley, M. *The Parietal Lobes* (Hafner, New York, 1953).
2. Lynch, J. C. *Behav. Brain Sci.* **3**, 484–534 (1980).
3. Andersen, R. A. in *Spatial Cognition: Brain Bases and Development* (eds Stiles-Davis, J., Kritchewsky, M. & Bellugi, U.) (University of Chicago Press, in the press).
4. Bock, O., Eckmiller, R. & Andersen, R. A. *Brain Res.* (in the press).
5. Andersen, R. A. in *Handbook of Physiology: The Nervous System V* 483–518 (eds Mountcastle, V. B., Plum, F. & Geiger, S. R. (1987).
6. Mountcastle, V. B., Lynch, J. C., Georgopoulos, A., Sakata, H. & Acuna, C. J. *Neurophysiology* **38**, 871–908 (1975).
7. Lynch, J. C., Mountcastle, V. B., Talbot, W. H. & Yin, T. C. T. *J. Neurophysiol.* **40**, 362–389 (1977).
8. Andersen, R. A. & Mountcastle, V. B. *J. Neurosci.* **3**, 532–548 (1983).
9. Andersen, R. A., Siegel, R. M. & Essick, G. K. *Expl Brain Res.* **67**, 316–322 (1987).
10. Andersen, R. A., Essick, G. K. & Siegel, R. M. *Science* **230**, 456–458 (1985).
11. Robinson, C. J. & Burton, H. *J. comp. Neurol.* **192**, 69–92 (1980).
12. Tanaka, K. *et al. J. Neurosci.* **6**, 134–144 (1986).
13. Saito, H. *et al. J. Neurosci.* **6**, 145–157 (1986).
14. Sakata, H., Shibutani, H., Ito, Y. & Tsurugai, K. *Expl Brain Res.* **61**, 658–663 (1986).
15. Gnadt, J. W. & Andersen, R. A. *Expl Brain Res.* (in the press).
16. Andersen, R. A. & Gnadt, J. W. in *Reviews in Oculomotor Research* Vol. 3 (eds Wurtz, R. & Goldberg, M.) (Elsevier, Amsterdam, in the press).
17. Andersen, R. A. in *Neurobiology of Neocortex. Dahlem Konferenzen* (eds Rakic, P. & Singer, W.) (Wiley, Chichester, in the press).
18. Andersen, R. A. & Zipser, D. *Can. J. Physiol. Pharmacol.* (in the press).
19. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* Vol. 1 (eds Rumelhart, D. E. & McClelland, J. L.) 318–362 (MIT, Cambridge, 1986).
20. Zipser, G. Institute for Cognitive Science, Report 8608 (UCSD, La Jolla, 1986).
21. Zipser, D. & Rabin, D. E. in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* Vol. 1 (eds Rumelhart, D. E. & McClelland, J. L.) 488–506 (MIT, Cambridge, 1986).
22. Ackley, D. H., Hinton, G. E. & Sejnowski, T. J. *Cognitive Sci.* **9**, 147–169 (1985).
23. Sejnowski, T. J., Kienker, P. K. & Hinton, G. E. *Physica* **22D**, 260–275 (1986).
24. Schlag, J., Schlag-Rey, M., Peck, C. K. & Joseph, J. *Expl Brain Res.* **40**, 170–184 (1980).
25. Peck, C. K., Schlag-Rey, M. & Schlag, J. *J. comp. Neurol.* **194**, 97–116 (1980).
26. Aicardi, G., Battaglini, P. P. & Galletti, G. *J. Physiol., Lond.* **390**, 271 (1987).

Spectrophotometry and polarimetry of the giant luminous arcs in the clusters 2244-02 and Abell 370

J. S. Miller & R. W. Goodrich

Lick Observatory, Board of Studies in Astronomy and Astrophysics, University of California, Santa Cruz, California 95064, USA

Lynds and Petrosian¹ have reported the discovery of very large luminous arcs in two clusters of galaxies, and there have been various conjectures² about the nature of these unusual features. Some of these are that they are: (1) images of distant objects formed by a gravitational lens generated by the visible clusters of galaxies; (2) regions of star formation stimulated by the passage of a shock front produced by a central (now extinguished) quasi-stellar object; (3) light echoes of a previously bright object at the centre; or (4) long arcs of stars tidally stripped from galaxies in the clusters. Observational data have been so limited that it has not been possible to establish any clearly favoured interpretation. Here we present new observations which restrict the range of possibilities, by eliminating (2), (3) and (4), even if they do not lead to a unique interpretation.

Using the Lick Observatory 3-m Shane reflector, we have obtained imaging polarimetry and spectrophotometry for the arc associated with the cluster 2244-02. We have also obtained a limited amount of spectrophotometric data for the arc associated with the cluster Abell 370. The cluster redshifts are 0.328 and 0.373 for 2244-02 and Abell 370, respectively. The polarimetry was obtained with a polarimeter consisting of a rotatable half-wave achromatic retarder and a pair of beam-splitting polarizing cubes, which produce two parallel beams of orthogonal polarizations. These result in two images of size $30'' \times 30''$. The detector (Texas Instruments) was a 800×800 three-phase CCD (charged-coupled device), which had been treated with NO gas and ultraviolet light to improve ultraviolet response and uniformity. The scale on the CCD is 0.72 arc s per pixel. Polarization modulation is achieved by taking separate exposures at four position angles of the retarder. (See ref. 3 for a description of the CCD, instrument and observing procedures.) Because the polarimeter is a two-beam device and the detector is extremely stable, our results are unaffected by atmospheric transparency changes from exposure to exposure. The polarimeter has been used regularly for more than 2 years and yields results whose accuracy depends only on photon-detection statistics. We used two configurations of the spectrograph: (1) an all-refracting system with a grism disperser that covered the wavelength region 4,400-7,200 Å with a resolution (two pixels sampling) of 7 Å; and (2) a reflecting system with good ultraviolet response that covered the region 3,000-6,000 Å with a resolution of 8 Å.

The imaging polarimetry data for 2244-02 were obtained on 1987 July 24 using an interference filter centred at 4,900 Å with full width at half maximum of ~ 800 Å. For the imaging polarimetry reductions, we subdivided the arc into four sectors of $\sim 20^\circ$ which were reduced independently. Four images (two from each of the two settings of the retarder) were used to calculate each of the two Stokes parameters Q and U , and the uncertainties were derived from the counting statistics. The accuracy was limited by the large contribution from the polarized sky emission. In the wavelength interval used, scattered light from nearby cities produces a sky polarization of $\sim 4\%$. Because the ratio of arc flux to sky flux was ~ 0.03 , we paid careful attention to sky subtraction and the errors introduced by the sky. The data for the four regions are consistent with zero polarization, and we are confident that the polarization in any of the regions measured must be $< 6\%$. If we take an average over all four regions, we derive an upper limit to polarization

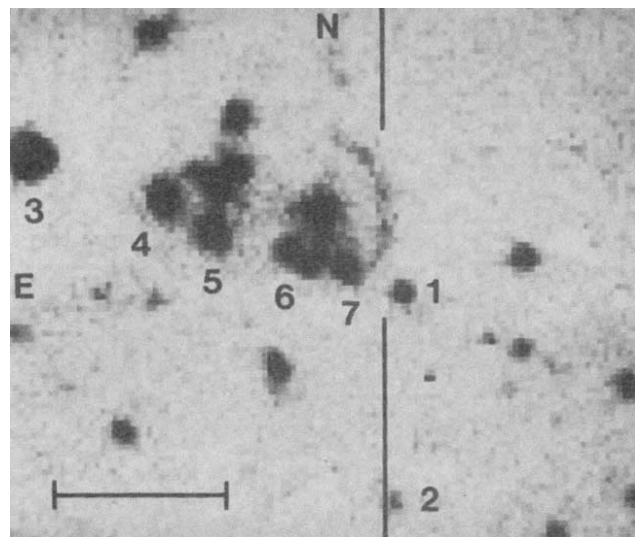


Fig. 1 CCD image of the cluster 2244-02. Spectra were obtained of galaxies 1-7; vertical line segments, slit position for the spectrophotometric observations of the arc; horizontal line at the lower left, 20 arc s on the sky.

of $\sim 4\%$. We estimate the uncertainty of the upper limits is $\sim 2\%$.

The violet-to-red spectral region of the 2244-02 arc was observed on 1987 August 24 and 25 for a total exposure time of 4 h. The ultraviolet-visual spectrum of this arc was recorded on August 26 with three hours of total exposure. For all spectra the spectrograph slit had dimensions $2.1'' \times 120''$ and was set at position angle 0° . Because the spectrograph also has an imaging mode, it was possible to position the slit accurately using direct images. Figure 1 shows one of these images and indicates the position of the slit. The arc associated with Abell 370 was observed with the ultraviolet-visual setup for a total exposure time of 160 min. The slit was set to position angle 108° and covered the arc from just to the east of galaxy SG to just to the north of galaxy C, where we used the designations given by Soucaill *et al.*⁴. The same detector was used for the spectroscopy and imaging as was used for the polarimetry. The spectrophotometric data were reduced to absolute fluxes using standard procedures and the standard stars calibrated by Stone⁵. Because of the extreme faintness of the arcs relative to the Mount Hamilton sky, great care was used in the sky subtraction. The sky background was modelled with a parabolic surface, and the spectra were extracted using the optimal extraction procedure of Horne⁶.

Figure 2 shows averages of all of our spectra of the arcs. In forming the averages we weighted the individual spectra according to the signal-to-noise ratio of the interval 4,500-5,500 Å. The noise in the final spectrum is highly wavelength-dependent because of wavelength-dependent sky contributions, atmospheric extinction, and system response. The noise is especially high near strong night-sky emission lines such as Na D $\lambda 5,893$ and [O I] $\lambda 6,300$. It is clear from Fig. 2 that the spectra show no strong emission or absorption features. For the arc associated with 2244-02, the continuum slope seems to change near 3,800 Å, with a decrease in flux towards the ultraviolet. Also, near the red end the flux appears to decrease. There are no obvious discontinuities in the spectrum. Soucaill *et al.*⁷ report that their spectrum of this arc in the range 4,400-7,200 Å is also flat with no easily identifiable features. The spectrum of the arc associated with Abell 370 also seems featureless. Although it also shows no obvious discontinuities, there are some departures from a straight line or line of slight curvature that are consistent with observational errors and the possible inclusion of a little

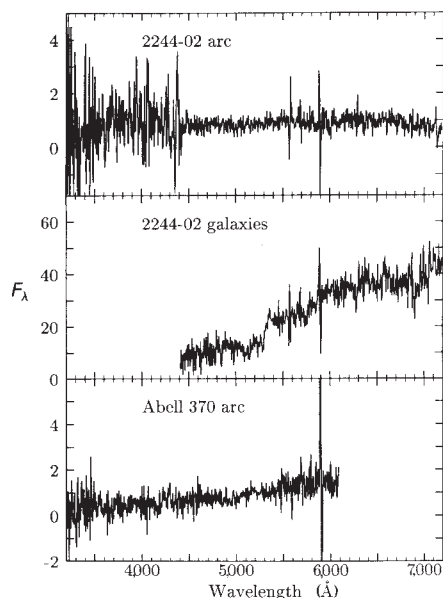


Fig. 2 Spectra of the arc and galaxies in the cluster 2244-02 and the arc in Abell 370. Middle panel, sum of the spectra of galaxies 1, 4, 5, 6 and 7. Flux scale in units of $10^{-17} \text{ erg cm}^{-2} \text{ s}^{-1} \text{ Å}^{-1}$.

light from galaxies near the arc, the latter more likely for $\lambda > 5,500 \text{ Å}$. These polarimetric and spectroscopic observations allow us to comment on various interpretations of the arc.

(1) Light echoes: Katz and Jackson (personal communication) have calculated the expected polarization for an arc produced by scattering of light from a central object. For all geometries except nearly perfect forward or back scattering, the polarization should be considerably higher than our upper limit. Further, the spectra of the arcs should be similar to that of a giant elliptical or quasar, which they are not. Therefore, we exclude this possibility.

(2) Tidally stripped tails of stars: The spectra of the arcs are incompatible with that expected from stars in the galaxies of the clusters. One would expect to see the continuum break at $4,000(1+z) \text{ Å} = 5,310 \text{ Å}$ for 2244-02 and $5,500 \text{ Å}$ for Abell 370, as well as other features typical of the late-type spectra of stars dominating the galaxies, and we do not. Our spectra of galaxies 1, 4, 5, 6 and 7 in 2244-02 are all very similar to one another, completely typical of elliptical galaxies and clearly different from the arc spectrum. Figure 2 also shows a combination of the spectra of these galaxies. Spectra for galaxies in Abell 370 are given in ref. 4.

(3) Star formation precipitated by an expanding shock front or cooling flow: In this case we would expect to see a much bluer continuum if the stars were spectral type early B or O and emission lines if the stars were O (unless star formation was extraordinarily efficient in converting gas to stars). Stars cooler than early O should show a pronounced Balmer discontinuity or a marked deficiency of light in the ultraviolet. None of the above is observed, so we also exclude this option.

(4) Synchrotron radiation in the cluster: This possibility is harder to exclude. The spectra could be consistent with a synchrotron mechanism, although some reddening would perhaps have to be invoked to account for small deviations from a power law. In this case the low polarization must be explained; this could be a result of disorder of the magnetic field lines. One would expect by analogy with known objects that if optical synchrotron emission were this luminous, then the arc should be a very luminous radio source and probably already catalogued. One would have to assume high optical depth in the radio to avoid high radio luminosity. Further, one would

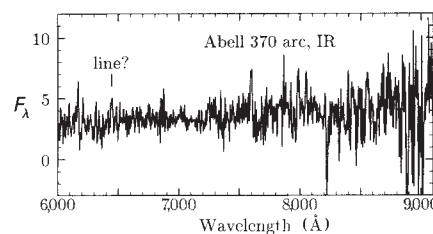


Fig. 3 Our 25 November 1987 result on the same scale as Fig. 2 (see text for details).

have to account for the giant arc shape for the emission, which is in strong contrast to what we know about the geometry of extragalactic radio sources. The synchrotron emission interpretation is unlikely, and a low or null detection in radio observations could strengthen this statement.

(5) Gravitational lens: The gravitational lens hypothesis is attractive because the explanations involving the arcs being parts of the clusters seem unlikely. But what kind of objects are being lensed? If they are normal galaxies or stars, the redshift must be high enough to have either the $4,000 \text{ Å}$ break or the Balmer discontinuity beyond about $6,500 \text{ Å}$ to make detection difficult, which implies redshift $z > 0.62-0.78$. However, Lyman- α and the Lyman discontinuity must also be at $\lambda \leq 3,600 \text{ Å}$, which requires $z \leq 2.0-3.0$. This is a large range, and it is possible that normal galaxies with perhaps some reddening could be the objects being lensed. Note that the slit position used for 2244-02 passed through galaxy 2 (Fig. 1), and the spectrum of this object shows a featureless, rather blue continuum, similar to that of the arc, plus two narrow emission lines which we identify with $[\text{O II}] \lambda 3,727$ and H_β at a redshift of 0.43. The results presented by Dressler and Gunn⁸ suggest that galaxies with a smooth blue continuum with no strong absorption features are not uncommon at faint magnitudes, so the objects being lensed could be one of these peculiar objects.

In conclusion, our data allow the elimination of several explanations for the giant arcs according to which they are physically associated with the clusters in which they appear. The gravitational lens hypothesis remains attractive, even though the nature of the object being lensed is unclear. We note that both 2244-02 and Abell 370 are clusters with two separate concentrations of galaxies, the arcs being on the outside of the overall distribution rather than between the two concentrations. Perhaps this is a coincidence, but preliminary calculations of gravitational lenses that produce arcs carried out by W. Burke (in preparation) suggest that having two mass concentrations is a good way to form a dominant arc-like image and hide other images formed by the lens.

After submission of this paper, Soucail *et al.*⁹ reported the detection of $[\text{O II}] \lambda 3,727$ emission and $\text{Mg II} \lambda 2,800$ absorption at a redshift of 0.724 in the arc associated with Abell 370. We obtained on 25 November 1987 an additional spectrum for this arc with a 2-h exposure. The spectrograph and slit position were the same as described above, and the interval from $6,000$ to $9,100 \text{ Å}$ was recorded; Fig. 3 shows the result on the same flux scale as Fig. 2. We see marginal evidence for an emission line at $6,443 \pm 4 \text{ Å}$, and inspection of the spectrum in Fig. 2 shows a feature at $4,852 \pm 4 \text{ Å}$ which could be identified as an absorption line. If we identify these features as $[\text{O II}]$ and Mg II , we derive $z = 0.733 \pm 0.003$, slightly different from their results. These are weak features in our data whose reality is open to question, but we can say that the spectrum does not show the $4,000 \text{ Å}$ break at this redshift or any other obvious spectral feature; apparent features near $7,280$ and $7,590 \text{ Å}$ are artefacts of the night-sky subtraction. These new results add support to the idea that the arc is the result of a gravitational lens, but the object being imaged is not similar to normal, nearby galaxies.

We thank Laura Kay for assistance in the observations and W. Burke for discussions. This research was supported in part by NSF.

Received 29 October; accepted 23 November 1987.

- Lynds, R. & Petrosian, V. *Bull. Am. astr. Soc.* **18**, 1014 (1986).
- Paczynski, B. *Nature* **325**, 572-573 (1987).
- Robinson, L. B. *et al. Opt. Engng* **26**, 795-805 (1987).
- Soucail, G., Fort, B., Mellier, Y. & Picat, J. P. *Astr. Astrophys.* **172**, L14-L16 (1987).
- Stone, R. P. S. *Astrophys. J.* **218**, 767-769 (1977).
- Horne, K. *Publs astr. Soc. Pacif.* **98**, 609-617 (1986).
- Soucail, G., Mellier, Y., Fort, B., Mathez, G. & d'Odorico, S. *IAU Circ. No. 4456* (1987).
- Dressler, A. & Gunn, J. E. *Astrophys. J.* **270**, 7-19 (1983).
- Soucail, G. *et al. IAU Circ. No. 4482* (1987).

Hyper-velocity and tidal stars from binaries disrupted by a massive Galactic black hole

J. G. Hills

Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico, USA

A close but newtonian encounter between a tightly bound binary and a $10^6 M_\odot$ black hole causes one binary component to become bound to the black hole and the other to be ejected at up to $4,000 \text{ km s}^{-1}$. The discovery of even one such hyper-velocity star coming from the Galactic centre would be nearly definitive evidence for a massive black hole. The new companion of the black hole has a high orbital velocity which increases further as its orbit shrinks by tidal dissipation. The gravitational energy released by the orbit shrinkage of a such a tidal star can be comparable to its total nuclear energy release.

The dynamics of gas in the Galactic centre (GC) suggest that it may have a $10^6 M_\odot$ black hole¹. Such a massive black hole would, by capturing the debris of stars that pass within its Roche radius, grow in less than 10^9 yr to $10^8 M_\odot$, which is the mass of black holes expected in luminous quasars^{2,3}. A GC black hole more massive than $4 \times 10^4 M_\odot$ is predicted to emit more non-stellar radiation than is observed from the GC, due to capture of stellar tidal debris². Although this is a severe problem, it does not totally rule out a $10^6 M_\odot$ GC black hole.

Simulations of close encounters between binaries and stellar intruders⁴ have shown that in exchange collisions, in which a massive intruder replaces one binary component, the average gravitational binding energy of the binary increases with the mass, M_1 , of the captured intruder. The increase is fivefold if M_1 is 100 times the mass of the binary components. I show here that this trend continues for encounters with much more massive intruders. The computations were done with the Shampine-Gordon⁵ variable-order, variable-step-size integrator. The mean fractional change in the energy of the system due to integration errors was 10^{-7} , so the high ejection velocities found in these simulations are not numerical artefacts. I used newtonian mechanics. Most encounters occurred well beyond the last stable circular orbit at three Schwarzschild radii.

In these simulations each binary component is of $1 M_\odot$ and the black hole masses are $M_{\text{bh}} = 10^4, 10^5, 10^6$ and $10^7 M_\odot$. The semi-major axis of the binary is $a_0 = 0.01 \text{ AU}$, except for two series of runs with $a_0 = 0.02$ and 0.1 AU for $M_{\text{bh}} = 10^6 M_\odot$.

Only 'hard' binaries, with orbital velocities larger than the mean velocity in the stellar system, survive encounters with fellow stars of approximately the same mass⁶⁻⁸. The maximum semi-major axes of hard binaries in galactic nuclei, where stellar velocities are typically $80\text{--}300 \text{ km s}^{-1}$, are in the range $a_0 = 0.01\text{--}0.1 \text{ AU}$. The velocity dispersion in the GC is also in this range⁹. The average mass density within 0.25 pc of the GC

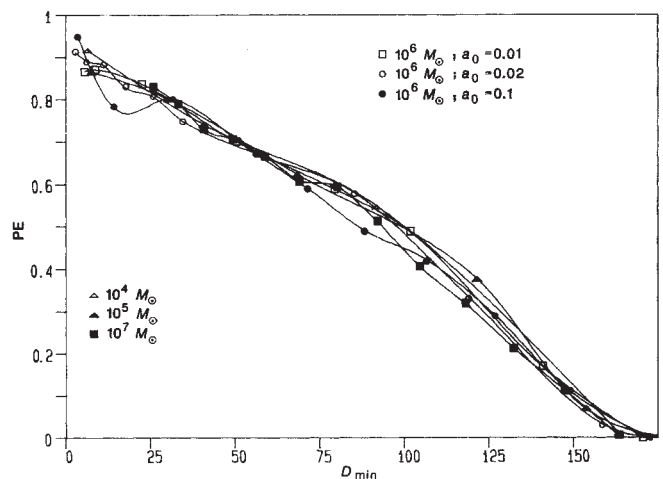


Fig. 1 Probability of an exchange collision in which the black hole replaces one of the original binary components, plotted as a function of the dimensionless closest approach parameter D_{min} of the binary to the black hole. The curves are labelled by the mass of the black hole and the semi-major axis of the binary orbit. The smallest D_{min} plotted for each of the two most massive black holes is just inside the last stable circular orbit.

exceeds $10^7 M_\odot \text{ pc}^{-3}$ (ref. 9), so close encounters between stars are common.

I simulated encounters having a wide range of impact parameters. At each impact parameter, 250 runs were made using different binary orientations randomly chosen with respect to the approaching intruder, which has a velocity at infinity of $V = 250 \text{ km s}^{-1}$. At this velocity, the total energy of the three-body system (binary plus black hole) is positive even for $a_0 = 0.01 \text{ AU}$ (by design, so the encounters are flybys). It is thus energetically possible for the encounter to dissociate the binary, but in almost every case the encounter left behind either the original binary or one in which the black hole had replaced one star. Because V is much less than either the velocity of the binary with respect to the black hole at closest approach or the recoil velocity of the ejected binary component, the results should not be sensitive to the chosen value of V .

The binary is 1,500 times its semi-major axis from the black hole at the beginning of each computation, far enough away for the tidal field of the black hole to be small but close enough so the binary comes to pericentre with the black hole after relatively few integration steps. I treat the stars as point masses, but they are broken apart if they pass within their Roche radius of the black hole. While tidal breakup is not a problem for neutron stars and white dwarfs, it is serious for main-sequence stars because of their relatively low densities. A $1 M_\odot$ main-sequence star is torn apart by a $10^6 M_\odot$ black hole if it passes within $R_{\text{min}} = 150 R_\odot = 0.70 \text{ AU}$ of the black hole (W. Benz and J. G. Hills, in preparation). Because lower main-sequence stars are denser than the Sun, they can pass significantly closer to the black hole before tidal breakup. A main-sequence star of $0.1 M_\odot$ has one-fourth the Roche radius of a $1 M_\odot$ star.

For a $1 M_\odot$ main-sequence star, the Roche radius exceeds the Schwarzschild radius if the black hole is less massive than $3 \times 10^8 M_\odot$ (ref. 2), so such a star is much more likely to be tidally torn apart by the black hole than to be swallowed whole.

Figure 1 shows the probability PE of an exchange collision as a function of the dimensionless closest approach parameter $D_{\text{min}} = (R_{\text{min}}/a_0)[2 M_{\text{bh}}/10^6(M_1 + M_2)]^{-1/3}$, where R_{min} is the closest approach of a binary with semi-major axis a_0 and component masses M_1 and M_2 to a black hole of mass M_{bh} . The factor in the square bracket compensates for the increase in the tidal interaction radius of the black hole as M_{bh} increases. Exchange collisions occur in nearly 90% of the closest encounters. We note that PE is nearly the same at a given value of

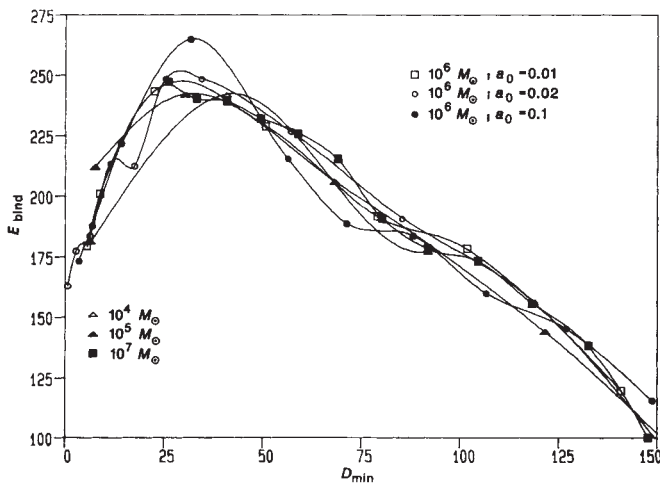


Fig. 2 The increase in the dimensionless binding-energy parameter of the binary as the result of an exchange collision with a black hole.

$D_{\min}$ for various values of M_{bh} and a_0 , so we can use Fig. 1 to find PE for arbitrary values of a_0 , $R_{\min}$ and M_{bh} .

Figure 2 shows E_{bind} , the normalized (dimensionless) increase in the average binding energy of those binaries which suffer exchange collisions. Here

$$E_{\text{bind}} = \left(\frac{\text{Increase in binding energy}}{\text{Original binding energy}} \right) \left(\frac{2 M_{\text{bh}}}{10^6 (M_1 + M_2)} \right)^{-1/3}$$

We note that the results are again nearly independent of a_0 and M_{bh} when given as a function of $D_{\min}$.

Figure 3 shows the ejection velocity parameter

$$V_{\text{ejection}} = V_{\infty} \left(\frac{2 M_{\text{bh}}}{10^6 (M_1 + M_2)} \right)^{-1/6} \times \left[\left(\frac{0.01 \text{ AU}}{a_0} \right) \left(\frac{M_1 + M_2}{2 M_{\odot}} \right) \right]^{-1/2}$$

for the ejected star in these exchange collisions. Here V_{∞} is the r.m.s. velocity at infinity of the ejected star as found in the numerical experiments. Figure 3 shows that V_{ejection} is well-behaved, so it is easy to determine V_{∞} for arbitrary values of M_{bh} and a_0 . For a black hole with $M_{\text{bh}} = 10^6 M_{\odot}$, r.m.s. ejection velocities at infinity are about $V_{\infty} = 4,000, 3,000$ and $1,400 \text{ km s}^{-1}$ for $a_0 = 0.01, 0.02$ and 0.1 AU respectively. These velocities greatly exceed that of any known star in the Galaxy, and the two higher ones are much larger than the escape velocity from the Galaxy.

The high ejection velocities occur because V_{break} , the centre-of-mass velocity of the binary at the point where the tidal field of the black hole pulls it apart, is much greater than the internal orbital velocity, V_{orb} , of the binary components. If each binary component experiences an average change in its velocity with respect to the black hole on the order of V_{orb} because of the perturbation of its companion, the corresponding change in its specific kinetic energy is $\delta KE \approx V_{\text{orb}} V_{\text{break}}$. For $M_{\text{bh}} = 10^6 M_{\odot}$, $V_{\text{break}} \approx 0.14c = 10^2 V_{\text{orb}}$ for $a_0 = 0.01 \text{ AU}$, so the change in the specific kinetic energy of the binary components with respect to the black hole is more than two orders of magnitude larger than their original orbital binding energy.

For a binary to have an exchange probability of 50% or more in an encounter with a $10^6 M_{\odot}$ black hole requires it to pass within about $100a_0$ of the black hole, or within $R_{\min} = 1$ and 10 AU for $a_0 = 0.01$ and 0.1 AU respectively. For a stellar mass density at the GC of $10^7 M_{\odot} \text{ pc}^{-3}$, a star passes within 1 AU of a $10^6 M_{\odot}$ black hole about every 100 yr and within 10 AU every $10 \text{ yr}^{2,3}$. (Here gravitational focusing is important, so the cross-section for passing within distance $R_{\min}$ of the black hole is

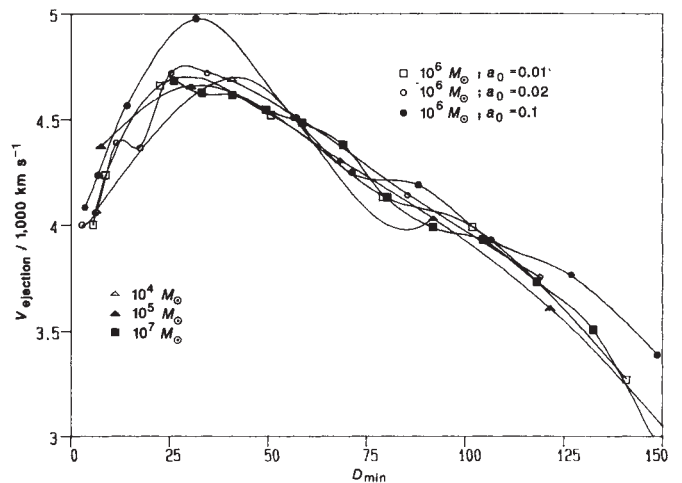


Fig. 3 The ejection velocity parameter of the unbound binary component after an exchange collision with the black hole.

proportional to $R_{\min}$ rather than to $R_{\min}^2$.) If 1% of the stars involved in these close encounters are binaries with $a_0 = 0.01 \text{ AU}$, and another 1% are binaries with $a_0 = 0.1 \text{ AU}$, then one of the closer binaries suffers an exchange collision about every 10^4 yr and one of the more distant binaries every 10^3 yr . (If the binary components are $1 M_{\odot}$ main-sequence stars, then about half of the encounters which are close enough to produce an exchange collision for $a_0 = 0.01 \text{ AU}$ lead to tidal breakup of the individual stars by the black hole, but the fraction of stars broken up is negligible in exchange collisions in which $a_0 = 0.1 \text{ AU}$.)

From Fig. 3 we see that the typical ejection velocity, V_{∞} , of the discarded binary component is $\sim 4,000 \text{ km s}^{-1}$ for $a_0 = 0.01 \text{ AU}$, and $1,400 \text{ km s}^{-1}$ for $a_0 = 0.1 \text{ AU}$, if $M_{\text{bh}} = 10^6 M_{\odot}$. Because it takes hyper-velocity stars with $V_{\infty} = 4,000 \text{ km s}^{-1}$ about $2 \times 10^6 \text{ yr}$ to reach the Sun's distance of $\sim 8 \text{ kpc}$ from the GC, there are about 200 of them between the GC and the solar circle if one is produced every 10^4 yr . Hyper-velocity stars with $V_{\infty} = 1,400 \text{ km s}^{-1}$ take at least $6 \times 10^6 \text{ yr}$ to traverse this distance even if we ignore their gravitational deceleration, so there are more than 6,000 of them inside the solar circle if one is produced every 10^3 yr .

The average space density of hyper-velocity stars decreases with the inverse square of their distance from the GC, with the closest one to the GC being within 40 pc for $V_{\infty} = 4,000 \text{ km s}^{-1}$ and within 1 pc for $V_{\infty} = 1,400 \text{ km s}^{-1}$. Proper motion survey fields near the GC but above the Galactic plane should be best for detecting hyper-velocity stars.

For $V_{\infty} = 4,000 \text{ km s}^{-1} = 800 \text{ AU yr}^{-1}$, the maximum proper motion of a hyper-velocity star at 8 kpc is $0.1 \text{ arc s yr}^{-1}$. A solar-luminosity star at this distance would be at magnitude 19 in the absence of interstellar extinction. Many other hyper-velocity stars will be closer, brighter, and have larger proper motions. If hyper-velocity stars exist, some may be included in existing proper motion surveys. The simplest explanation for their not being reported is that they do not exist, which implies that the mass of the GC black hole is much less than $10^6 M_{\odot}$. Alternatively, observers may have disbelieved the outlandish velocities implied by a facile interpretation of their proper motions and assumed these stars to be peculiar and much closer to us than they actually are. Hyper-velocity stars should also have large radial velocities which are hard to rationalize away. But radial velocities are much more expensive to determine than proper motions, so relatively few proper motion stars have known radial velocities.

Hyper-velocity stars may well be peculiar compared to stars in the solar neighborhood. Because they originate in the core of the Galaxy, their metal abundances may actually be several times greater than solar, and because they reach the solar

distance from the GC in much less than the solar Kelvin time, they may not have returned to the main sequence if they were strongly perturbed by the tidal field of the black hole. A strong tidal perturbation would also cause them to be in rapid rotation (W. Benz and J. G. Hills, in preparation).

All active galactic nuclei (AGNs) eject hyper-velocity stars which become intergalactic tramps. The expansion of the universe gradually slows them down with respect to the Hubble flow until they are captured by galaxies. Hyper-velocity stars from AGNs may constitute a minor impurity of metal-rich stars in the Galactic halo.

If hyper-velocity stars are not found, it may still be possible to retain the hypothesis of a massive GC black hole by arguing that my estimate of the binary frequency in the GC is too optimistic, that most GC stars are less massive, so fainter, than the Sun, or that exchange collisions between binaries and stellar remnants (white dwarfs, neutron stars, and small black holes) have allowed these hard-to-detect objects to replace so many main-sequence stars in GC binaries that few main-sequence stars are left in them¹⁰. But the detection of even one hyper-velocity star coming from the GC would be nearly definitive proof of its having a massive black hole.

After an exchange collision with a $10^6 M_{\odot}$ black hole, the semi-major axis of the binary is typically 50 AU for $a_0 = 0.01$ AU and 500 AU for $a_0 = 0.1$ AU. The r.m.s. orbital velocity, V_{orbital} , of the star about the black hole is nearly the same as the ejection velocity, V_{∞} , of its former companion, or about 4,000 and 1,400 km s^{-1} respectively for these two binaries.

The velocity of an unbound star which collides with one of these bound stars exceeds the local parabolic speed which is, on average, about $\sqrt{2} V_{\text{orbital}}$, so the two stars collide at an average speed of about $\sqrt{3} V_{\text{orbital}} = 6,900$ and $2,400 \text{ km s}^{-1}$ for the two cases. The minimum collision velocity required to break up two equally massive main-sequence stars is only about 2.3 times the escape velocity from their surface, or about $1,400 \text{ km s}^{-1}$ (ref. 11). Most of the debris from the colliding stars should remain bound to the black hole and eventually be accreted by it.

The orbit of the bound star around the black hole is very eccentric, but tidal energy dissipation at pericentre may be able to circularize it before the star is destroyed in a stellar collision. Because of angular momentum conservation, the final semi-major axis of the circularized orbit will be about $a_f = 2R_{\text{min}}$. Many tidal stars can accumulate in these close orbits because the high velocities in them make them very 'stiff', so it is very difficult for stars in other bound orbits to perturb them gravitationally.

The tidal energy dissipated in circularizing the orbits of stars captured in an exchange collision with a $10^6 M_{\odot}$ black hole is ~ 0.1 – 1% of their rest-mass energy, or 10^{51-52} erg for a $1 M_{\odot}$ star. (Comparable energy is released in one final supernova-class cataclysm when a tidal star in a nearly circularized orbit collides with an unbound star.) The radius of the captured tidal star increases, eventually following its Hayashi track, until its luminosity matches its rate of tidal energy dissipation. If each tidal star emits 10^{51} erg before its destruction in a collision and one is captured every 10^4 yr, the total integrated luminosity of the tidal stars orbiting the GC black hole is $\sim 10^6 L_{\odot}$. The maximum luminosity of an individual star is its Eddington value, so at least 30 solar-mass stars are needed to produce the integrated luminosity. The integrated spectrum of these stars should mimic that of a giant with its absorption lines washed out by the extreme differential Doppler motion resulting from the high orbital velocities (0.03 – $0.1 c$), of the individual tidal giants. The GC black hole may have accumulated a large number of these low-mass tidal giants in orbits with semi-major axes, a_i , on the order of few times R_{min} . It is tempting to identify a_i with radius of the GC radio source, ~ 10 AU. Stellar winds and Roche overflow from the tidal giants may terminate any accretion disk beyond a_i .

Stellar remnants such as neutron stars, small black holes, and

white dwarfs do not suffer appreciable tidal dissipation, so they tend to remain in their original post-exchange orbits. The GC black hole can accumulate a large number of these stellar remnants in orbits with semi-major axes of 50 – 500 AU and orbital velocities of $1,000$ – $4,000 \text{ km s}^{-1}$. They are 'meat grinders' which help provide fuel for their master, the central black hole, by tearing material from main-sequence stars and giants that pass near them.

Received 18 August; accepted 18 November 1987.

1. Oort, J. H. A. *Rev. Astr. Astrophys.* **15**, 295–362 (1977).
2. Hills, J. G. *Nature* **254**, 295–298 (1975).
3. Hills, J. G. *Mon. Nat. R. astr. Soc. Soc.* **182**, 517–536 (1978).
4. Hills, J. G. & Fullerton, L. W. *Astr. J.* **85**, 1281–1291 (1980).
5. Shampine, L. F. & Gordon, M. K. *Computer Solution of Ordinary Differential Equations: The Initial Value Problem* (Freeman, San Francisco, 1975).
6. Aarseth, S. J. & Hills, J. G. *Astr. Astrophys.* **21**, 255–263 (1972).
7. Heggie, D. C. *Mon. Nat. R. astr. Soc.* **173**, 729–788 (1975).
8. Hills, J. G. *Astr. J.* **80**, 809–825 (1975).
9. Sellgren, K., Hall, D. N. B., Kleinmann, S. G. & Scoville, N. Z. *Astrophys. J.* **317**, 881–891 (1987).
10. Hills, J. G. *Mon. Nat. R. astr. Soc.* **175**, 1p–4p (1976).
11. Benz, W. & Hills, J. G. *Astrophys. J.* **323**, 614–628 (1987).

Laboratory simulation of Jupiter's Great Red Spot

Joël Sommeria*, Steven D. Meyers & Harry L. Swinney

Center for Nonlinear Dynamics and the Department of Physics,
The University of Texas, Austin, Texas 78712, USA

Isolated large stable vortices have long been observed in the jovian atmosphere and more recently on Saturn. The existence of such stable vortices in strongly turbulent planetary atmospheres is a challenging problem in fluid mechanics. In a numerical simulation Marcus¹ found that a single stable vortex developed for a wide variety of conditions in a turbulent shear flow in a rotating annulus. To test this we conducted an experiment on a rotating annulus filled with fluid pumped in the radial direction. The annulus rotates rigidly (there is no differential rotation), but the action of the Coriolis force on the radially pumped fluid produces a counter-rotating jet. Coherent vortices spontaneously form in this turbulent jet, and for a wide range of rotation and pumping rates the flow evolves until only one large vortex remains.

Figure 1 shows the annular tank used in the experiments. The tank is rotated rapidly to achieve an essentially geostrophic flow (a two-dimensional flow in which the Coriolis force is balanced by pressure gradients²). In such a flow the inertial force is small compared with the Coriolis force, that is, the Rossby number is small. In our system the Rossby number is ~ 0.1 , as estimated from the ratio of the characteristic time for the inertial force (the turnover time for a vortex) to the characteristic time for the Coriolis force (the rotation period of the tank).

On Jupiter the Great Red Spot and other persistent coherent vortices lie in latitudinal zones of strong shear³. Marcus's numerical simulation indicates that in a quasi-geostrophic flow the existence of a strong shear can lead to the formation of a single persistent coherent vortex. In our experiment there is a strong shear on the outer edge of the counter-rotating jet (see Fig. 2).

Dissipation appears to play little role in the dynamics of the jovian atmosphere⁴, and the simulation by Marcus was done for an inviscid fluid. Our experiment is unusual in that the dynamical timescales of interest are typically an order of magnitude shorter than the viscous timescale. For example, the Ekman spin-down time t_E is typically 20 s while the vortex turnover time is only 2 s. (Measurements of t_E agree well with the calculated value, $t_E = h_0/2(\Omega\nu)^{1/2}$, where ν is the kinematic viscosity, Ω is the angular velocity of the tank, and h_0 is the mean depth

* Permanent address: Madylam, ENSHMG, BP 95, 38402 St Martin D'Hères Cedex, France.

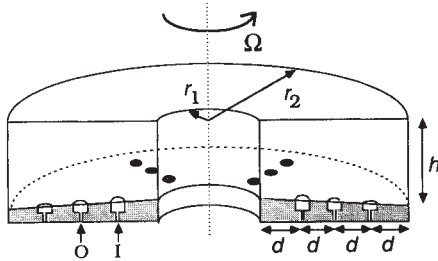


Fig. 1 The apparatus: $r_1 = 10.8$ cm, $r_2 = 43.2$ cm, $d = 8.1$ cm, $h(r_1) = 17.1$ cm and $h(r_2) = 20.3$ cm. The total flow rate F ranged from 15 to 370 cm³ s⁻¹, and the rotation rate Ω ranged from 0 to 25 rad s⁻¹. A counter-rotating jet was produced by the action of the Coriolis force on the fluid (water) pumped into the tank through a ring of six inlets (I) at $r = 18.9$ cm, and from the tank through a ring of six outlets (O) at $r = 27.0$ cm (the ring of ports at $r = 35.1$ cm was not used). The effect of this forcing on the boundary layers was minimized by using 500 small holes (0.08-cm diameter) instead of a single hole at each inlet and outlet.

of the tank.) The weak dissipation is balanced on average by the pumping of the fluid. The Reynolds number for the bulk flow is large (typically 5×10^4), but the Reynolds number for the thin boundary layers is small enough so that these layers are laminar.

In a planetary atmosphere the Coriolis force varies with latitude (the beta effect). This is mimicked in our tank by a radially sloping bottom (slope $s = -0.1$). The transparent cover of the tank is flat. These considerations of geostrophic balance, Ekman damping and the beta effect lead to the following equations of motion for the vorticity, $\omega = (\nabla \times \mathbf{u})_z$, in our system^{1,2}:

$$(\partial/\partial t + \mathbf{u} \cdot \nabla)(\omega + \beta r) = \zeta - \omega/t_E, \quad \nabla \cdot \mathbf{u} = 0 \quad (1)$$

where $\mathbf{u}$ is the velocity, $\zeta = 2\Omega w/h_0$ is the rate of vorticity production due to the forcing (with w the vertical velocity at an inlet or outlet), r is the radial coordinate, and $\beta = 2\Omega s/h_0$ is the coefficient of the beta effect. Both terms on the right-hand side of equation (1) are small; hence the potential vorticity of a fluid parcel, given here by $\omega + \beta r$ (see refs 1, 2 and 18) is to a first approximation conserved as it is advected. Also, the horizontal divergent flow due to the pumping is very small ($\sim 1\%$) compared with the nondivergent azimuthal flow generated by the strong Coriolis force.

Evidence for a strong counter-rotating jet is presented in Fig. 2, which shows profiles of the mean azimuthal velocity U , the vorticity $\omega_b = dU/dr + U/r$, and the shear $\sigma = dU/dr - U/r$. For small pumping rates the jet is confined between the inlets and outlets, while for large pumping rates the jet fills the annulus. Figure 2 is for a case of moderate pumping, where the jet, broadened by the turbulent mixing of vorticity, is wider than the separation between the inlets and outlets. Cyclonic vorticity (same sign as Ω) and anticyclonic vorticity (opposite sign) are produced at the same rate, but the shear in Fig. 2 is everywhere cyclonic, as in a single shear layer. (The anticyclonic vorticity on the inner side of the jet corresponds mostly to solid body rotation without shear.) The asymmetry between cyclonic and anticyclonic shear is a consequence of the jet curvature, and would disappear in the limit of small jet width (relative to the radius).

The dependence of the width $2L$ and maximum velocity U_0 of the jet on the pumping and rotation rates (F and Ω) can be estimated from a straightforward physical argument: the torque arising from the action of the Coriolis force on the radially pumped fluid, $\int 2\Omega \langle u_r \rangle (2\pi r^2 h_0) dr = 4\pi \Omega \langle u_r \rangle h_0 r_m^2 d$, is balanced by the torque arising from the dissipation in the boundary layers, $\int r U(r) (2\pi r h_0) dr/t_E = 2\pi h_0 r_m^2 U_0 L/t_E$, where r_m is the radial position of the jet, d is the separation between the sources and sinks, and $\langle u_r \rangle = F/2\pi r_m h_0$ is the radial velocity averaged in time and along a perimeter. Equating these torques yields $U_0 L =$

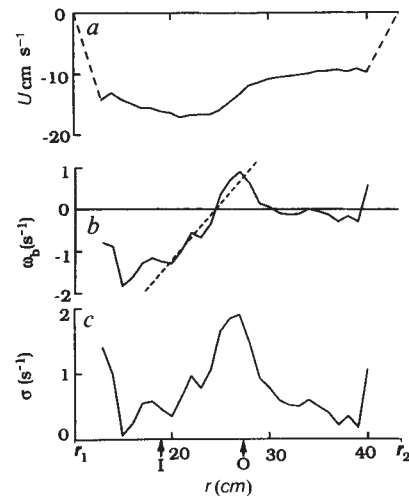


Fig. 2 *a*, The radial dependence of the azimuthal velocity $U(r)$, averaged in time and azimuth in a region far from any vortex. (The rotation rate was 25 rad s⁻¹; the total flow rate, 140 cm³ s⁻¹). The velocity field was determined from analyses of photographs of particle streaks (see Fig. 3). *b*, The vorticity $\omega_b(r)$, determined from $U(r)$. The straight line is βr , corresponding to uniform potential vorticity (see text). *c*, The shear $\sigma(r)$, determined from $U(r)$. The spot is centred at the position of maximum shear. (The vertical arrows on the abscissa indicate the radial positions of the inlets (I) and outlets (O).)

$2\Omega \langle u_r \rangle t_E d$. (The same result for $U_0 L$ can be obtained by multiplying equation (1) by r^2 and integrating over the annulus.)

Another relationship between L and U_0 can be obtained from the observation that the maximum slope of the vorticity, $d\omega_b/dr$, is approximately equal to β , as Fig. 2*b* shows. This result, which holds over a wide range of Ω and F values, can be interpreted as a consequence of the mixing of potential vorticity by the turbulent flow. As $d\omega_b/dr \approx U_0/L^2$, we have $U_0/L^2 = \beta$. Combining this expression for U_0/L^2 and the one for $U_0 L$ above we obtain

$$U_0 = c_1 (F\Omega)^{2/3} \quad (2)$$

$$L = c_2 F^{1/3} \Omega^{-1/6} \quad (3)$$

where $c_1 = (sd^2/2\pi^2 \nu h_0 r_m^2)^{1/3}$ and $c_2 = (dh_0/4\pi s r_m \nu^{1/2})^{1/3}$. Our measurements of the maximum velocity of the jet agree with equation (2) within 20% or better.

When the pumping rate is increased from low values with Ω held fixed, the jet develops (at $F \approx 25$ cm³ s⁻¹) an instability that results in the formation of a stable ring of five cyclonic vortices in the cyclonic shear on the outer side of the jet (see Fig. 3*a*). With a further increase in the pumping rate, a critical value is reached at which two vortices merge and a stable ring of four vortices forms (see Fig. 3*b*). Further increases in pumping rate lead to similar transitions to three and then to two stable vortices. At each of these steps the distance between the vortices increases more than the vortex size, and the position of the vortices fluctuates more and more as the flow becomes more turbulent with the increased forcing.

Finally, when the pumping rate is increased beyond ~ 100 cm³ s⁻¹, we reach a regime with a single robust cyclonic vortex, see Fig. 3*e*. This regime persists to the highest flow rate achievable with our pump, 370 cm³ s⁻¹. The single vortex moves with an average velocity equal to the local background flow and remains near the radial position of the sinks. Anticyclonic vortices also form in the turbulent flow, but they are always small and short-lived, disappearing in about one vortex turnover time.

The dynamics can be characterized in terms of two independent dimensionless parameters, chosen as

$$L^* = L/r_m \quad (4)$$

$$T^* = t_E L/U_0 \quad (5)$$

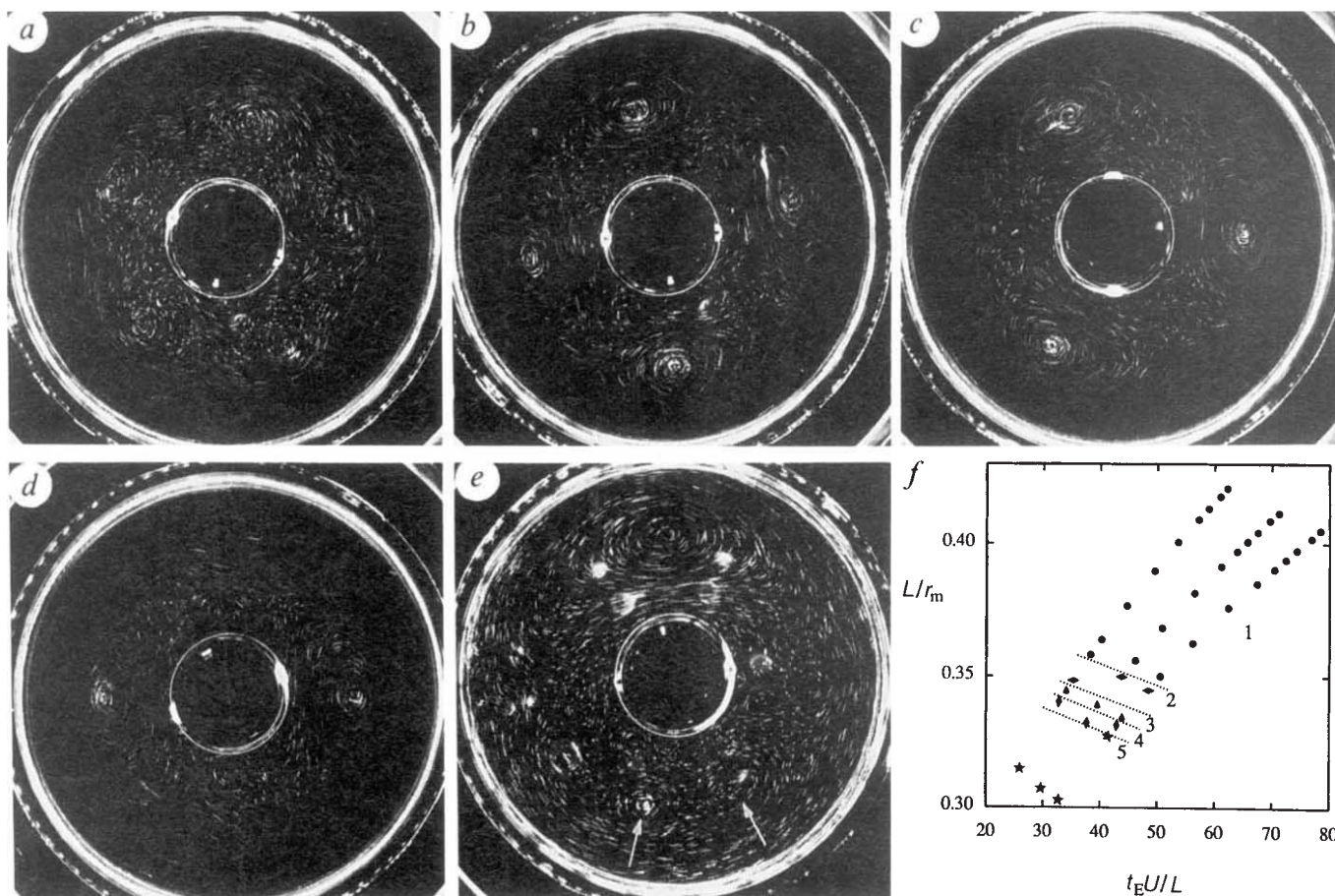


Fig. 3 Photographs of particle streaks for flows with different numbers n of persistent vortices: *a*, $n=5$; *b*, $n=4$; *c*, $n=3$; *d*, $n=2$; *e*, $n=1$. (The flow rates were, respectively, 25, 55, 66, 77, 185 cm³ s⁻¹; tank rotation rates, 3, 3, 3, 3, 4 Hz; and photograph exposure times, 0.5, 0.5, 0.5, 0.5, 0.12 s.) The camera was rotating in the co-rotating frame of the average motion of the vortices, which in (*e*) was 0.8 rad s⁻¹ slower than the rotation rate of the tank. Note that in (*e*), in addition to the main vortex, two smaller vortex pairs (indicated by arrows) can be seen to be in the process of merging. *f*, A regime diagram showing the dependence of the number of vortices as a function of the dimensionless jet size L^* and the ratio T^* of the Ekman friction time t_E to the vortex turnover time L/U (see text). Different symbols are used to distinguish regimes with different numbers of stable vortices, and dashed lines are shown to guide the eye in separating these regimes. The multiple vortex regimes are found only for small T^* where frictional effects are more important.

L^* is the dimensionless jet width and T^* is the ratio of the Ekman friction time to an eddy turnover time L/U_0 . Substituting the expressions for L and U_0 from equations (2) and (3), we have

$$L^* = (c_2/r_m) F^{1/3} \Omega^{-1/6} \quad (6)$$

$$T^* = (h_0 c_1 / 2 c_2 \nu^{1/2}) (F \Omega)^{1/3} \quad (7)$$

Figure 3*f* shows how the dynamics depends on these dimensionless parameters. The regime with a single persistent vortex is the dominant regime—multiple vortices are obtained only for small L^* and T^* .

When dye is injected inside a large vortex, it stays there for a long time with little indication of mixing with the outside fluid. When dye is injected outside a vortex, it is quickly mixed with the background turbulent flow, but does not penetrate the vortex, even after several minutes (see Fig. 4). The penetration of the dye is inhibited by a slow outward radial flow in the vortex, due to a recirculating current generated in the boundary layers. This circulation, called Ekman pumping², replaces the fluid inside the vortex in a typical time t_E . We can deduce from this that the turbulent diffusion across the vortex boundary into the vortex has a timescale t_E . The loss of dye from the vortex core occurs mainly during the merging of vortices.

In the regime with a single stable vortex, small cyclonic vortices are continually formed over the sinks, but they always merge with one another or with the dominant vortex. Figure 4

shows the vortex-merging process. The process is rapid, typically occurring in one vortex turnover time. In regimes with two or more stable vortices, mergings are not observed except when there is a transition between different regimes.

The dynamics leading to vortex mergers is easily understood. Because vortices move with the speed of the background flow, two vortices at different radii will move at different speeds, ultimately approaching one another. Then, if a secondary vortex is slightly downstream of the main vortex, it will be pushed outward by the flow of the main vortex and the background shear flow will carry the secondary vortex towards the main vortex. A similar process occurs when the secondary vortex is initially upstream.

When a secondary vortex is generated on the opposite side of the tank from the main vortex, but at the same radius, it can develop for a fairly long time, gaining in strength until it is comparable in size to the main vortex. However, this symmetrical arrangement is unstable—a fluctuation in the radial position of a vortex will result in the vortices approaching one another, and then they will quickly merge. When the forcing is increased, the fluctuations increase and the coexistence of vortices of comparable size becomes less and less frequent.

The main vortex grows by merger with smaller vortices until it fills the zone of strong shear. Then, in further mergers the excess vorticity feeds the background flow and often re-coalesces

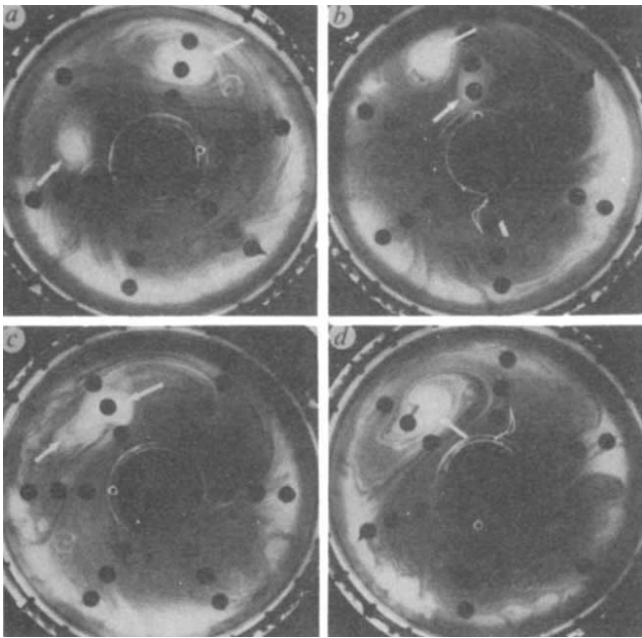


Fig. 4 The merging of two vortices (each indicated by an arrow) is illustrated in this sequence of photographs (taken at time intervals of 2 s) of the flow visualized with injected dye. The dye, injected outside of the vortices, does not reach the vortex cores. The 18 black dots are inlets and outlets in the bottom of the tank (see Fig. 1). The system is in the regime with a single stable vortex; the parameters are the same as in Fig. 3e.

into new small vortices; this can be observed in photographs of dye-containing vortices, which release part of the dye in the merging process.

The growth and merging of vortices observed in the experiments is very similar to that found in the numerical simulations of Marcus¹—merging occurs in a time comparable with the vortex turnover time, and a vortex grows until it fills a shear zone. In the simulation the potential vorticity is uniform throughout the annulus, while in the experiment the region of uniform potential vorticity is limited to the azimuthal band in which the vortex is confined, and radiation of Rossby waves² is possible outside this band. In both the simulation and the experiment the vortices have about the same shape and magnitude relative to the background cyclonic shear flow, and cyclonic vortices are stable and anticyclonic vortices unstable in this flow. (In the simulation, anticyclonic vortices were found to be stable in an anticyclonic shear, but our experiment was not designed to produce a well-mixed anticyclonic shear; for example, if the direction of pumping in the experiment is reversed, a wavy coherent jet forms and there is little mixing.) Thus, although it is not possible to compare quantitatively the laboratory observations for a viscous fluid with the numerical simulation for an inviscid fluid, the qualitative properties of the vortices in the simulation are strikingly similar to those observed in the laboratory.

The vortices that we observe in the single vortex regime also have many properties in common with jovian vortices^{3,5,6}, including the following: vortices in the same latitudinal zone approach one another and quickly merge; they are usually located at a latitude of maximum shear; they have a vorticity with the same sign as the background shear (just as on Jupiter where anticyclonic spots occur in zones of anticyclonic shear and cyclonic barges occur in zones of cyclonic shear); they drift with the velocity of the background flow; the mean excess vorticity¹ of a spot is comparable to the background shear; and the spot length-to-width ratio is about two.

Our system, as a model of Jupiter, neglects the vertical struc-

ture of the jovian atmosphere, simplifying it to a single layer of uniform density and thickness. A refined model would include the interaction of the layer of strong eddy activity with layers above and below it. A measure of the vertical stratification is a horizontal length scale, called the Rossby radius of deformation¹, over which the surface of the fluid layers deform, intensifying or weakening lines of vorticity. The distribution of vorticity within a vortex thus depends on the Rossby radius, and in fact the maximum vorticity for jovian vortices is in an annular ring while our vortices have maximum vorticity at their centre. However, except for that difference, jovian vortex dynamics does not seem to be dramatically modified by the deformation—this view is supported by the observation that jovian spots range in size considerably ($\sim 10^3$ – 10^4 km), compared to the radius of deformation (estimated to be 1,000–5,000 km^{1,6,7}).

In our experiment the Ekman pumping leads to a weak downwelling in the upper half of the centre of a vortex and a return flow on the outside of the vortex; similarly, there is an upwelling in the lower half of a vortex centre. Similar circulations have been observed in jovian vortices^{7–9}. As in the experiment, such circulations could arise from a frictional mechanism, although the friction of upper atmospheric layers for jovian vortices would be different from that of the solid boundary in the experiment.

The forcing for our system is different from that on Jupiter, but our pumping could have an effect similar to the convective plumes penetrating into the jovian atmosphere from below. Direct observations of the flow and of weak temperature gradients in the jovian atmosphere indicate that the behaviour is governed largely by quasi-horizontal inertial effects⁹; therefore, the details of the forcing mechanism may not actually be important in determining the eddy patterns, even though forcing is essential for maintaining the energy and momentum of the flow.

Previous experiments have also been conducted on flows driven by sources and sinks in rotating tanks^{10,11}, but those experiments did not have a strong shear. We believe that the shear is the main ingredient for the formation of stable vortices, not only in our study but also in previous experiments aimed at modelling long-lived jovian vortices^{12,13}. Two-dimensional shear layers have a strong and general tendency to develop robust vortices of the same sign as the shear, and in fact rings of steady vortices have been observed in previous studies of annular shear flows^{12–15}. However, previous observations of single isolated vortices have been limited to narrow ranges of experimental parameters.

In an experiment by Antipov *et al.*¹² shear was produced by differential rotation of annular rings in the bottom of a parabolic-shaped rotating tank. The free upper surface deformed in response to the pressure field induced by the flow. Vorticity sheets, similar to those above our rings of sinks and sources, were generated at the boundaries between the rings and the tank bottom. A stable anticyclonic vortex was observed to form and was interpreted to correspond to the propagation of a Rossby soliton^{16,17}. With cyclonic shear Antipov *et al.* obtained a ring of cyclonic vortices, similar to our observations for small pumping rates. They observed no merging of vortices, which is consistent with their interpretation of the observed vortex as a Rossby soliton but contrasts with our observations and those for Jupiter. The Antipov *et al.* system is similar to ours, but there are two differences: (1) the scale of forcing, as discussed below, and (2) the dissipation, which was larger in their 1-cm-thick layer (for which $T^* \approx 5$) than in our experiment (where $25 < T^* < 80$).

In another experiment Read and Hide¹³ obtained a stable anticyclonic vortex (paired with a weaker cyclonic vortex) in a rotating annulus that was heated internally and cooled on the vertical sidewalls. The vertical structure of the flow was found to be consistent with the structure proposed for Jupiter. However, a single anticyclonic vortex was observed only in a

narrow laminar regime close to the transition from axisymmetric to non-axisymmetric flow, far from the regime of geostrophic turbulence. In the experiments of both Hide and Read and Antipov *et al.* the forcing was axisymmetric, which contrasts with the localized forcing produced by our pumping through inlets and outlets. The small scale of our forcing could be significant because such forcing is generally a necessary condition for strong turbulence in a quasi-two-dimensional flow (energy transfers towards small scales are inhibited).

Our experiment demonstrates that for a wide range of conditions a permanent coherent vortex, like the long-lived coherent spots on Jupiter, can emerge spontaneously from a turbulent, incompressible, quasi-geostrophic flow. It is not necessary to invoke stratification or deformation of the fluid layer to explain the formation of a stable spot. The dynamical behaviour is dominated by two-dimensional inertia and the beta effect; frictional effects are secondary, as on Jupiter.

Our experiments and the simulation of Marcus¹ suggest that long-lived vortices should form in planetary zones of strong shear and uniform potential vorticity. Recent estimates¹⁸ indicate that the potential vorticity is fairly uniform for the zonal flows that contain the Great Red Spot and the White Oval BC, but further measurements are needed for other latitudes where long-lived vortices are located.

We thank Philip Marcus for suggesting this experiment. This research is supported by the NSF Program in Fluid Mechanics and Hydraulics, the Office of Naval Research Nonlinear Dynamics Program, and the Exxon Education Foundation. J.S. is supported in part by the Centre National de la Recherche Scientifique.

Received 16 June, accepted 11 December (1987).

- Marcus, P. S. *Nature* **331**, 693–696 (1988).
- Pedlosky, J. *Geophysical Fluid Dynamics* (Springer, New York, 1979).
- Beebe, R. F. & Hockey, T. A. *Icarus* **67**, 96–105 (1986).
- MacLow, M. M. & Ingersoll, A. P. *Icarus* **65**, 353–369 (1986).
- Mitchell, J. L., Beebe, R. F., Ingersoll, A. P. & Garneau, G. W. *J. geophys. Res.* **86**, 8751–8757 (1981).
- Williams, G. P. *Adv. Geophys.* **28A**, 381–427 (1985).
- Ingersoll, A. P. & Cuong P. G. *J. atmos. Sci.* **38**, 2067–2076 (1981).
- Hatzes, A., Wenkert, D. D., Ingersoll, A. P. & Danielson, G. E. *J. geophys. Res.* **86**, 8754–8749 (1981).
- Flasar, F. M. *Icarus* **65**, 280–303 (1986).
- Colin de Verdier, A. *Geophys. Astrophys. Fluid Dyn.* **15**, 213–251 (1980).
- McEwan, A. D., Thompson, R. O. R. Y. & Plumb, R. A. *J. Fluid Mech.* **9**, 655–672 (1980).
- Antipov, S. V., Nezlin, M. V., Snezhkin, E. N. & Trubnikov, A. S. *Nature* **323**, 238–240 (1986).
- Read, P. L. & Hide, R. *Nature* **308**, 45–48 (1984).
- Hide, R. & Titman, C. W. *J. Fluid Mech.* **29**, 39–60 (1967).
- Rabaud, M. & Couder, Y. *J. Fluid Mech.* **136**, 291–319 (1983).
- Maxworthy, T. & Redekopp, L. G. *Icarus* **29**, 261–271 (1976).
- Petviashvili, V. I. *Soviet Astr. Lett.* **9**, 137–138 (1983).
- Dowling, T. E. & Ingersoll, A. P. *J. atmos. Sci.* **45** (1988).

Numerical simulation of Jupiter's Great Red Spot

Philip S. Marcus

Department of Mechanical Engineering, University of California at Berkeley, Berkeley, California 94720, USA

Jupiter's Great Red Spot is viewed as a vortex that arises naturally from the equations of motion of the jovian atmosphere. Here I solve numerically the equations governing fluid motion in a model of the jovian atmosphere for a variety of initial conditions. Large spots of vorticity form spontaneously in chaotic azimuthal flows and are stable if the vorticity of the spots has the same sign as the shear of the surrounding azimuthal flow. The Great Red Spot is compared with these solutions and a new prediction of its vertical structure is made.

Following Ingersoll and Cuong¹, I use a two-layer model in which the Great Red Spot lies in a shallow layer overlying a

deep azimuthal flow. In this model the shallow upper layer does not influence the denser flow beneath it, but the deep azimuthal flow affects the upper layer by determining its vertical depth. The velocity v of the upper layer is approximated as two-dimensional (due to the rapid rotation of the planet) and is determined by potential vorticity ω_p conservation:

$$\left[\frac{\partial}{\partial t} + (v \cdot \nabla) \right] \omega_p \equiv \frac{D\omega_p}{Dt} = 0 \quad (1)$$

We use the quasi-geostrophic β -plane approximation² for ω_p of the upper layer:

$$\omega_p(r, \phi, t) = \omega(r, \phi, t) - \beta r - \frac{\psi(r, \phi, t) - \bar{\psi}(r)}{L_R^2} \quad (2)$$

where β is the gradient of the Coriolis force, ϕ is the longitude, r is the latitude of Jupiter multiplied by its radius, $L_R \equiv (\sqrt{gH\Delta\rho/\rho})/f$ is the Rossby deformation radius (a measure of the vertical stratification), f is twice Jupiter's angular velocity, $\Delta\rho$ is the difference in density between the lower and upper layers, and ρ , H , g , $\psi(r, \phi, t)$ and $\omega(r, \phi, t) \equiv \nabla^2\psi(r, \phi, t)$ are the density, mean vertical depth, gravitational acceleration, stream function and vorticity of the upper layer. The azimuthal velocity of the lower layer is given by the stream function $\bar{\psi}(r)$. The physics responsible for the deep flow (convection, latent heat release, radiation and differential heating, for example) occurs on timescales that are much longer^{1,3} (10^8 s) than the dynamical or turn-around time of the spot ($\sim 10^6$ s). (In the five months between the fly-bys of Voyagers 1 and 2, there was no appreciable change in Jupiter's azimuthal zones⁴.) Therefore we treat $\bar{\psi}(r)$ as a time-independent, but unknown, parameter that we vary to obtain different numerical solutions. Here, equations (1) and (2) are solved numerically with spectral methods⁵ and with impermeable boundaries at $r = R_1$ and R_2 . We use cylindrical, rather than the usual cartesian coordinates in equations (1) and (2) to allow a direct comparison with the cylindrical laboratory experiments of Sommeria *et al.*⁶.

I adopt the hypothesis that the jovian zone containing the Great Red Spot is an approximately axisymmetric, azimuthal flow with nearly uniform ω_p that contains isolated patches of non-uniformity (including the red spot). Although the hypothesis cannot be tested directly because $\bar{\psi}(r)$ and L_R are unknown, it is consistent with their known bounds¹. The motivation for this hypothesis is that numerical simulations have shown that fluids that obey equation (1) and that are mixed by external stirring tend to produce flows with nearly uniform ω_p over much of their domains^{7,8}. This evolution to homogeneous ω_p has also been observed in the laboratory⁶. Our numerical simulations of equations (1) and (2) in an annular geometry with no stirring also show that large regions with homogeneous ω_p form when the initial velocity is random but constrained so that its azimuthal average, $1/(2\pi) \int_0^{2\pi} v_\phi(r, \phi, t=0) d\phi$, is a linearly unstable solution of equation (1)⁹. (All axisymmetric, azimuthal flows are exact solutions of equation (1).)

Typically, the nearly axisymmetric, azimuthal flows with uniform ω_p produced in our simulations and in the experiments of Sommeria *et al.* contain a few, small, isolated vortices where the value of ω_p differs substantially from that of the surrounding azimuthal flow. Because the evolution of a general class of initial flows into nearly uniform ω_p flows containing isolated vortices is discussed elsewhere^{6,9}, each numerical simulation presented here begins with an exactly axisymmetric, azimuthal, uniform ω_p flow superposed with a finite number of spots of vorticity. We define $\tilde{v}_\phi(r, t)$ as the velocity of the uniform ω_p component of the flow, $\sigma \equiv r\partial(\tilde{v}_\phi/r)/\partial r$ as its shear, and ω_e as the difference between the potential vorticity of a spot of vorticity and the ω_p of $\tilde{v}_\phi(r, t)$. Note that $\tilde{v}_\phi(r, t)$ is a neutrally stable solution of equation (1); it has no exponentially growing or decaying eigenmodes and cannot propagate Rossby waves⁹.

First, I show that spots with ω_e the same sign as σ are stable and those with opposite sign are pushed to the radial boundaries

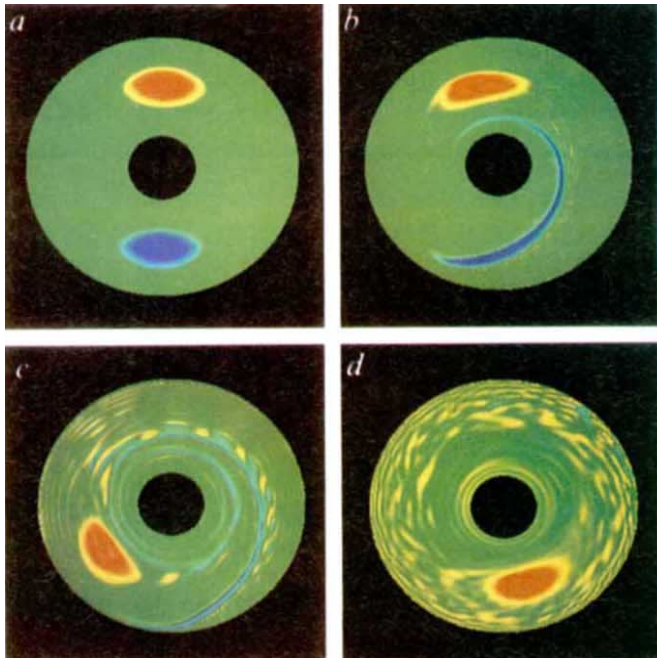


Fig. 1 The ω_e of two spots superposed with an axisymmetric, azimuthal, uniform ω_p flow with negative shear. The colour scale follows the order of the spectrum: $\omega_e < 0$ is red, $\omega_e = 0$ is green, and $\omega_e > 0$ is blue. The time is in units of the turn-around time of the initial spots (defined in the text). *a*, $t = 0.0$; *b*, $t = 0.50$; *c*, $t = 1.5$; *d*, $t = 4.8$.

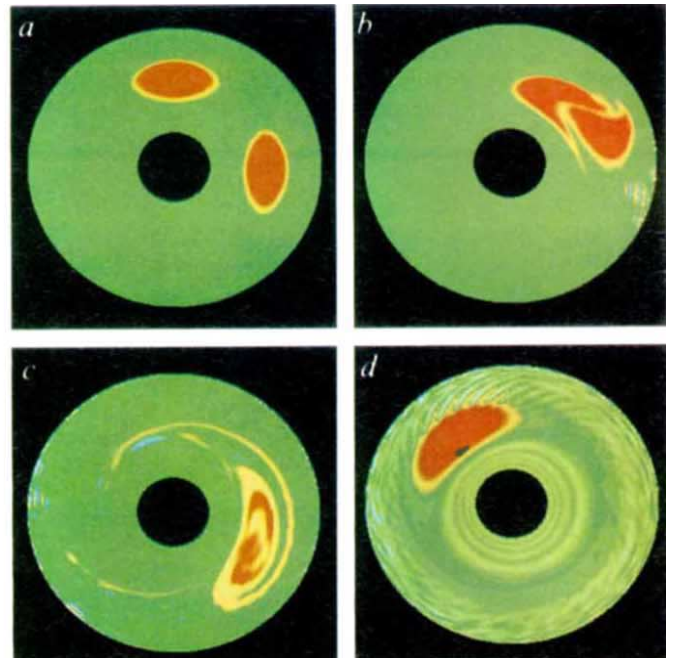


Fig. 2 Two stable spots initially superposed with an axisymmetric, azimuthal, uniform ω_p flow with negative shear. *a*, $t = 0.0$; *b*, $t = 1.3$; *c*, $t = 2.6$; *d*, $t = 28$.

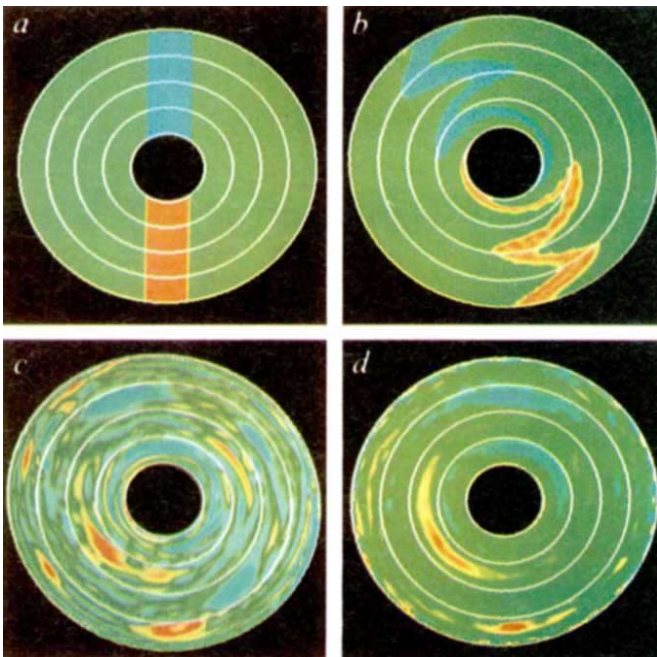


Fig. 3 Rectangular spots of positive and negative ω_e superposed initially with an axisymmetric, azimuthal, uniform ω_p flow with a sinusoidally varying shear. The white circles mark the radii where the shear changes sign. (These are also the extrema of the angular velocity of the azimuthal flow.) *a*, $t = 0.0$; *b*, $t = 0.057$; *c*, $t = 1.1$; *d*, $t = 11$.

of the azimuthal flow and fragment. Figure 1 shows a flow consisting initially of two spots of vorticity of equal strength and with approximately uniform ω_e but with opposite sign. The azimuthal flow superposed initially with the spots is $\tilde{v}_\phi(r, t = 0) = \beta r^2/3 + C/r$, where C is a constant equal to $\frac{1}{6}$. The flow has $\beta = 1$, $R_1 = 0.25$, $R_2 = 1.0$ and $L_R \rightarrow \infty$. The shear of the initial $\tilde{v}_\phi$ is $\sigma(r, t = 0) = \beta r/3 - 2C/r^2$, and is negative throughout the entire domain of the flow. Both spots have $|\omega_e| \approx 0.9$ (so $|\sigma/\omega_e| \approx 0.7$).

In Fig. 1*b* the red spot, with ω_e of the same sign as σ , is stable and becomes more elliptical, but the blue spot stretches into a spiral with its inner and outer edges pushed to the inner and outer boundaries, respectively. The spiral is Kelvin-Helmholtz unstable^{5,9}; it fragments and breaks apart (Fig. 1*c*). Finally, Fig. 1*d* shows an isolated red spot superposed on a chaotic (as determined by broad peaks in its power spectrum) azimuthal flow containing many small yellow and blue filaments. If the azimuthal flow is averaged over scales large compared to the

filaments but small compared with the red spot, it has uniform ω_p . The spot is advected with the velocity of the surrounding azimuthal flow. Although it continually loses and gains ω_e , it is statistically steady. The behaviour shown in Fig. 1 is representative for a wide range of values of the parameters β , ψ , L_R and C , including the case $\beta = \psi(r) = 0$ (neither the gradient of the Coriolis force nor an underlying topography are essential), as long as $10 > |\sigma/\omega_e| > 0.1$ and $L_R \geq L_H$, where L_H is the characteristic length over which ω_e changes (not necessarily the same as the size of a spot—see below). When parameters are varied so that $|\sigma/\omega_e|$ or L_R decreases, the stable spots become more elongated in shape⁹. When $|\sigma/\omega_e| > 10$ they are so elongated that they can be Kelvin-Helmholtz unstable. When $|\sigma/\omega_e| < 0.1$ the spots behave as if there were no surrounding azimuthal flow present. Flows with $\sigma > 0$ have stable blue spots.

The merger of two stable spots is shown in Fig. 2. The flow parameters and the initial $\tilde{v}_\phi$ are the same as in Fig. 1, but both initial spots in Fig. 2a have $\omega_e < 0$. The spots are initially towards each other by the differential rotation of $\tilde{v}_\phi$. As the spots approach, they distort, shed some ω_e (Fig. 2b), and merge, trapping weakly rotational fluid (yellow-coloured) at the vortex centre (Fig. 2c). The yellow streamer trailing from the vortex is Kelvin-Helmholtz unstable⁹ and becomes wavy. Figure 2d shows a statistically steady spot, superposed on a chaotic azimuthal flow. Yellow filaments torn from the spots make the azimuthal flow chaotic, but the azimuthal flow has uniform ω_p when averaged over length scales larger than the size of the filaments. The timescale for mergers is the turn-around time of a spot, $\tau \approx 4\pi/\omega_e$. By varying the parameters and initial conditions, we have found that the size of the critical impact parameter (initial separation in r) necessary for two spots to merge increases approximately linearly with σ/ω_e and approximately equals the sum of the radial axes of the two merging spots. A derivation of the critical impact parameter based on the assumption that energy in the chaotic component of the flow is maximized was given earlier⁹. We have found circumstances where two spots do not merge if surrounded by a laminar azimuthal flow, but do merge if surrounded by a chaotic flow with small inhomogeneities in ω_p . A possible explanation is that the dynamics of two isolated spots is constrained to conserve energy and momentum. Small inhomogeneities in the azimuthal flow can exchange energy and momentum with the spots, relax the constraint, and thereby allow mergers.

Notice that between Fig. 2c and d the spot has expelled the weakly rotational, yellow fluid from its centre. We have never found a stable spot with a local minimum of ω_e at its centre. Flows that initially have $\tilde{v}_\phi$ superposed with many small spots, each with a different value of ω_e , can produce one large spot by mergers. Moreover, if the flow is so chaotic that the small spots are well-mixed with each other before and during the merging, the core of the large spot will have approximately uniform ω_e .

Figure 3 shows the dynamics of spots in a uniform ω_p flow where the shear of $\tilde{v}_\phi$ does not have a constant sign. Initially, $\omega(r, t=0) = 3 \sin[4\pi(r-R_1)/(R_2-R_1)]/r$ and $\tilde{v}_\phi(r, t=0) = r \int_{R_1}^r \sigma(r', t=0)/r' dr' - 0.46r$. The parameters $\tilde{\psi}(r)$ and L_R^2 were chosen such that $\tilde{\psi}(r)/L_R^2 = \beta r - (1/r)(\partial/\partial r)[r\tilde{v}_\phi(r, t=0)]$ and $\psi(r, t)/L_R^2 \rightarrow 0$ (this limit corresponds to the ratio of the fluid density in the upper to lower layers approaching zero). The white circles show the radii where $\sigma = 0$. If we consider the annular regions between the circles as 'zones', then the σ of the four zones alternates in sign with the innermost zone having $\sigma > 0$. Initially, red and blue rectangular spots are superposed with $\tilde{v}_\phi$, but these distort due to the differential rotation of $\tilde{v}_\phi$ (Fig. 3b). The distortion is at a maximum where $\tilde{v}_\phi/r$ has an extremum (at the white circles). The large spots quickly fragment, and pieces with ω_e of the opposite sign of the local $\sigma(r, t)$ are expelled from the zones (Fig. 3c). The small pieces of ω_e coalesce and form one large spot in each zone (Fig. 3d). The

spots alternate in colour from zone to zone and are highly elongated because $\sigma/\omega_e \approx 5$.

Figure 3 shows several properties common to all of our simulations: several zones with opposite shear can co-exist; stability of spots does not depend on the proximity of horizontal boundaries; the sign of the local σ determines which spots of ω_e are stable and which are ejected from a zone; spots grow by amalgamating other spots until they either overflow their own zone (see Fig. 3d) or have amalgamated all available ω_e (which occurs when the initial spots are smaller than those in Fig. 3a). Flows with spots that do not fill their entire zones tend to move to the radial location where $\sigma(r)$ have a local extremum, in agreement with the laboratory study of Sommeria *et al.*⁶. (An isolated spot can change its radial location only if it exchanges energy and momentum with another spot or with the small filaments of ω_e in the surrounding azimuthal flow.)

Although the dynamics of spots is similar throughout the range $L_R \geq L_H$, the vorticity and velocity distributions inside the spots vary qualitatively with the value of L_R : a uniform ω_e spot with finite L_R concentrates most of its vorticity at the edge of the spot and has a non-rotating centre. In contrast, a spot with uniform ω_e and with $L_R \rightarrow \infty$ has approximately uniform vorticity. This can be demonstrated analytically for the case $\beta = \sigma = \tilde{v}_\phi = 0$ where the spots are circular. The vorticity in a circular spot with $L_R \rightarrow \infty$ is exactly uniform, but with finite L_R , the vorticity is proportional to $I_0(\hat{r}/L_R)$, where $\hat{r}$ is the distance from the centre of the spot and I_0 is a modified Bessel function that increases exponentially in $\hat{r}$. The vorticity drops abruptly by ω_e at the edge of the spot, and decreases exponentially (proportional to the Bessel function $K_0(\hat{r}/L_R)$) outside the spot.

The parameters of Jupiter's Great Red Spot, $\sigma/\omega_e \approx 1$ and $L_H/L_R \approx 1$, are within the range of values presented here, and the vorticity distribution and shape of the red spot agree qualitatively with our simulations. The dynamics of the red spot looks similar to our numerical simulations, especially the amalgamation of smaller spots, which often leaves trailing wisps of ω_e downstream of the red spot¹⁰. The red spot, like the flow in Fig. 3d, fills its zone (in the latitudinal direction) and like all of our simulated spots, does not drift with respect to $\tilde{v}_\phi$ except for the chaotic buffeting it gets from the small filaments embedded in $\tilde{v}_\phi$.¹¹ (Note that the slow 90-day zonal oscillation of the red spot¹³ cannot be produced by our simulations.) There are other long-lived vortices on Jupiter with dynamics similar to the red spot¹¹ and our simulations. Although some rotate clockwise and others anti-clockwise, all of these spots rotate in the same direction as the local shear, which is consistent with our numerical simulations.

Because the velocity of the red spot increases exponentially outward (with $L_H \approx 2,200$ km) and has a non-rotating centre with most of its vorticity concentrated at its periphery, it is possible that it has uniform ω_e . This suggests that it was created by a chaotic amalgamation of smaller spots. To see if ω_e is uniform in the core of the red spot and to find the value of L_R at the spot, consider the variation of ω_e as a function of longitude (but at fixed latitude). Equivalently, consider the difference in ω_e between the points (r_c, ϕ) and (r_c, ϕ_c) where ϕ_c and r_c are the longitude and latitude (multiplied by the radius of Jupiter) of the centre of the spot. From equation (2) this difference is

$$\begin{aligned} \omega_e(r_c, \phi) - \omega_e(r_c, \phi_c) &= \omega(r_c, \phi) - \omega(r_c, \phi_c) - \frac{\psi(r_c, \phi) - \psi(r_c, \phi_c)}{L_R^2} \\ &= \omega(r_c, \phi) - \omega(r_c, \phi_c) + \frac{r_c \int_{\phi_c}^{\phi} v_r(r_c, \phi') d\phi'}{L_R^2} \end{aligned} \quad (3)$$

where I have assumed that L_R is a function of r but not ϕ , and I have used $\psi(r_c, \phi) - \psi(r_c, \phi_c) = -r_c \int_{\phi_c}^{\phi} v_r(r_c, \phi') d\phi'$. The advantage of equation (3) over equation (2) is that it does not contain the unknown function $\psi(r)$. If the left-hand side of equation (3) is zero, the spot has uniform ω_e . Substituting the

observed values¹² of $v_r(r_c, \phi)$ and $\omega(r_c, \phi)$ into the right-hand side, we find that the left-hand side is ~ 0 when $L_R \approx 1,300$ km. The fit is particularly good for distances $< 9,000$ km from the centre of the red spot. (The semi-major axis of the red spot is $\sim 13,000$ km.) This calculation of L_R is the first to be made based on the dynamics of the red spot, and it is consistent with the estimates based on the thermal structure of the atmosphere³. This calculation suggests that the red spot has a large, 18,000-km diameter core with a nearly uniform potential vorticity, that the size of the spot is much greater than L_R , and that $L_R \approx L_H$.

The picture here is that the red spot is embedded in a chaotic but approximately uniform ω_p zone. It formed by amalgamating the small patches of ω_e that continually form on Jupiter due to small-scale processes (first suggested by Ingersoll¹), grew in size until it overflowed its zone, and is currently in statistical equilibrium—continually exchanging ω_e with the surrounding flow. Although my two-layer model of the jovian atmosphere is simplistic, it produces robust, long-lived vortices that are dynamically similar to the red spot. Future calculations will need to include a more realistic vertical structure. My assumption that the zone surrounding the red spot has a nearly uniform ω_p will also have to be tested. The spots of vorticity produced by our model are stable if they have the same sense of rotation as the zonal shear, can exist in chaotic flows, and tend to coalesce. The spots with finite L_R are like the jovian red spot in having

quiet interiors with vorticity mostly at their edges. The spots do not require the proximity of horizontal boundaries, nor finite values of β , L_R or ψ for stability. The values of these parameters do not strongly influence the dynamics of spots, but they do determine the sign and strength of $\tilde{v}_\phi(r, t)$ and the vorticity distribution and shape of spots.

I thank H. L. Swinney and J. Sommeria for discussions and insight and B. Scott for the colour figures. The work was sponsored by NSF, the Office of Naval Research and the Lawrence Livermore National Laboratory.

Received 25 June; accepted 11 December 1987.

1. Ingersoll, A. P. & Cuong, P. G. *J. Atmos. Sci.* **38**, 2067–2076 (1981).
2. Pedlosky, J. *Geophysical Fluid Dynamics* (Springer, New York, 1979).
3. Conrath, B. J., Flasar, F. M., Pirraglia, J. A., Gierasch, P. J., Hunt, G. E. *J. geophys. Res.* **86**, 8769–8775 (1981).
4. Limaye, S. S. *Icarus* **65**, 335–352 (1986).
5. Marcus, P. S. in *Numerical Analysis* (eds Griffiths, D. F. & Watson, G. A.) 125–139 (Longman Scientific and Technical, Harlow, 1986).
6. Sommeria, J., Meyers, S. D., Swinney, H. L. *Nature* **331**, 689–693 (1988).
7. Rhines, P. B. & Young, W. R. *J. Fluid Mech.* **122**, 347–367 (1982).
8. Williams, G. P. *J. Atmos. Sci.* **36**, 932–968 (1979).
9. Marcus, P. S. *J. Fluid Mech.* (submitted).
10. Low, M.-M., M. & Ingersoll, A. P. *Icarus* **65**, 353–369 (1986).
11. Beebe, R. F. & Hockey, T. A. *Icarus* **67**, 96–105 (1986).
12. Mitchell, J. L., Beebe, R. F., Ingersoll, A. P. & Garneau, G. W. *J. geophys. Res.* **86**, 8751–8757 (1981).
13. Smith, B. A. & Hunt, G. E. in *Jupiter* (ed. Gehrels, T.) 564–585 (University of Arizona Press, Tucson, 1976).

Direct identification of superconducting regions in an inhomogeneous specimen by electron shadow microscopy

Yang Cui-Ying & J. W. Steeds

Physics Department, University of Bristol, Bristol BS8 1TL, UK

Scanning and transmission electron microscopy have played a major part in evaluating the microstructure of the new high-temperature superconductors. Complex fabrication routes have led to a wide variety of phases in the samples produced, and the high-temperature superconductor itself has a complex microstructure on which its superconducting properties are dependent. Finally, the process of preparing specimens for examination by electron microscopy can itself lead to degradation of the superconducting properties. As a result it is desirable to establish directly that the regions studied are indeed superconducting, and to relate the superconducting properties to the local microstructure. Here we describe a convenient method for performing such an evaluation. One possible method, which relies on image contrast in a scanning electron microscope, was recently proposed¹. Another, which we discuss here, involves the use of a long established, if little used, technique of electron shadow microscopy².

The technique is closely related to Schlieren microscopy. Flux exclusion from superconducting regions leads to local changes in the magnetic field which cause Lorentz deflection of the high-energy electrons in a transmission electron microscope. The focusing effects of the complicated form of the resultant magnetic fields give rise to the well-known classes of catastrophes in optics which are now well-understood³. In the simplest case, cycloidal shadows are produced⁴, and numerical calculations of the shadow patterns produced by superconducting lead cylinder were performed⁵.

In order to pursue these experiments, we decided to abandon the fracture chip method of specimen preparation mostly used in previous transmission electron microscopy experiments on the new superconductors. Our main aim was to relate microscopic observations to macroscopic specimens, and for this purpose we decided to ion-thin sections of the material prepared. On fracturing, the location of a fragment in the original specimen

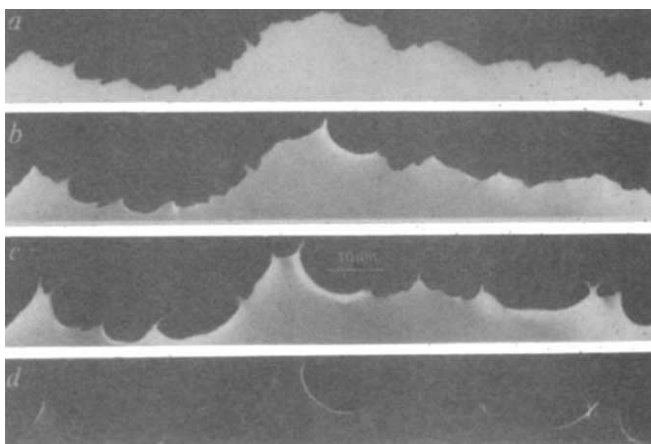


Fig. 1 Shadow image of the edge of an ion-thinned sample at the approximate temperatures a, 90 K; b, 74 K; c, 35 K; d, 35 K. Short exposure was employed to reveal details of the caustic figures.

is lost; moreover the technique is generally selective and fails to pick out all the components of a multiphase preparation. However, ion-thinning of samples is known to introduce damage⁶ and considerable care had to be exercised in obtaining satisfactory samples.

It was also important to have a double-tilting liquid helium cooled specimen holder available for the experiments reported here. Even though the superconducting transition temperature of $\text{YBa}_2\text{Cu}_3\text{O}_x$ is well above that of liquid nitrogen, it is unlikely that fracture chips reach this temperature in liquid nitrogen cooled holders. Cooling by conduction via the tortuous path required to reach the fracture chip is unlikely to be very effective. In addition, the thin specimens, with aspect ratios which are typically not less than 100:1, were mounted perpendicular to a magnetic field of ~ 1.5 T. The resulting demagnetizing fields will have caused flux penetration into the sample.

Some initial work was performed with fracture chip specimens to identify unambiguously the $\text{YBa}_2\text{Cu}_3\text{O}_x$ phase in our preparations (see, for example, refs 7 and 8). After a while it was possible to pick these out by their twinned microstructure,

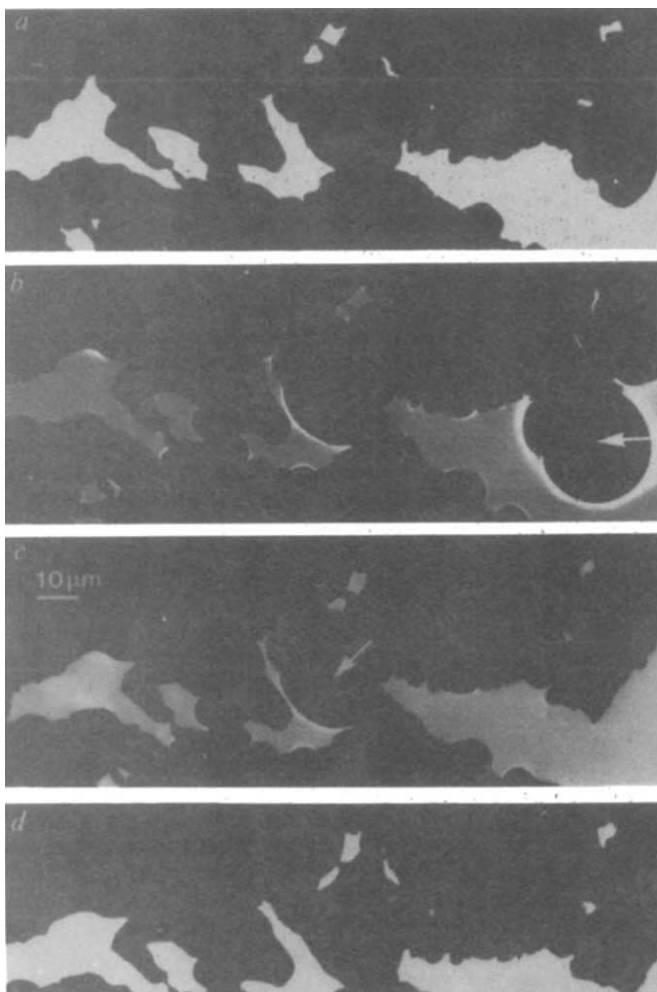


Fig. 2 Shadow image of the edge of an ion-thinned sample in a transmission electron microscope with a maximum operating voltage of 300 kV. The images were obtained under the following conditions: *a*, at the outset, temperature 90 K; *b*, at 30 K before electron irradiation; *c*, at 30 K after irradiation of the area arrowed at the right-hand side of *b*; *d*, at 30 K after irradiation of the area arrowed at the centre of *c*.

angular shape and energy-dispersive X-ray fingerprint. Samples were then prepared by bombardment of liquid-nitrogen-cooled thin sections of material with 5-kV argon ions in an Ion Tech FAB 406P ion thinner. Some fracturing and particle loss invariably occurred. Figure 1 shows a sequence of shadow images of the edges of a specimen, as the sample was cooled from room temperature to ~35 K. The caustic shapes first appeared when the thermocouple mounted on the heat exchanger recorded 90 K (Fig. 1*a*), and grew in scale as the temperature dropped (Fig. 1*b* and *c*). Figure 1*d* is a shorter exposure of Fig. 1*c* which reveals details of the caustic formation more clearly. On warming, the caustics reduced in amplitude and vanished at ~90 K. After performing this experiment the temperature dependence of the magnetic susceptibility of the sample was investigated. A sharp reduction in the magnetic susceptibility was found, corresponding to the superconducting transition at ~90 K.

Although these results seem at first sight conclusive evidence of a Meissner effect and possible charging effects associated with granular superconductivity, it is necessary to eliminate other possibilities. It is quite likely that other parts of the specimen may be semiconducting in nature. It is less likely, but not impossible, that the resistance of such regions might increase to such a level at ~90 K and even more at lower temperatures,

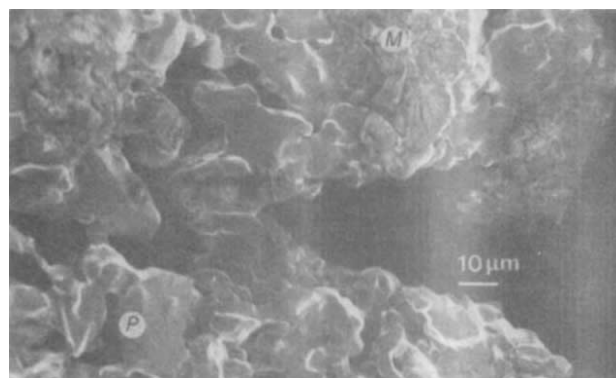


Fig. 3 Scanning electron micrograph of the same area of specimen as that illustrated in Fig. 2 after the irradiation experiment.

so that considerable electrostatic charging occurred leading to similar caustic effects. To examine this possibility, the following experiment was performed. An ion-thinned sample was first taken through the cooling sequence discussed above (Fig. 2*a*, *b*), and regions giving rise to strong caustic formation were identified. One of these was then subjected to a short period of intense irradiation by a focused 300-kV electron beam at room temperature. The sample was then once again cooled and this region then failed to produce a caustic figure, Fig. 2*c*. The experiment was then repeated on another region, to produce the low temperature result shown in Fig. 2*d*. The loss of caustic formation after electron irradiation strongly suggests that superconductivity was responsible for the effect.

The areas giving caustic figures at low temperature were too thick even for transparency to 300-kV electrons. The sample was therefore transferred to a scanning electron microscope to investigate the area giving the caustics. A complex microstructure was observed with smooth-faceted particles (P) embedded in a fibrous matrix (M) (Fig. 3). Energy dispersive X-ray (EDX) microanalysis was performed in the various different regions of the specimen, and the resulting spectra were used to identify the $\text{YBa}_2\text{Cu}_3\text{O}_x$ phase in scanning electron micrographs such as Fig. 3. The smooth angular particles were those which produced caustic figures and gave EDX spectra consistent with $\text{YBa}_2\text{Cu}_3\text{O}_x$. The regions damaged in the electron irradiation experiments described above had greatly reduced yttrium levels, although the barium/copper ratio remained that of the $\text{YBa}_2\text{Cu}_3\text{O}_x$ phase, and quite different from the adjoining matrix which was much richer in barium.

These experiments indicate that it is indeed possible to locate local regions of superconducting phase in an inhomogeneous specimen by shadow electron microscopy. The failure to observe the effects in electron-transparent regions (≥ 300 nm) may be the result of flux penetration due to the very high local fields which would occur. However, experiments were performed with very low excitation of the objective lens, and as the caustics still failed to appear in the greatly reduced magnetic field, it might be concluded that there were dead layers associated with the free surfaces in our samples, and that thin regions do not superconduct. If this is the case it would eliminate the possibility of direct visualization of vortex lattices by Lorentz deflection of electrons penetrating thin specimens⁹. There is independent evidence of normal layers of these superconductors¹⁰.

As the regions responsible for caustic generation were opaque, there was necessarily some ambiguity about the nature of these regions. It is evident from Fig. 3 and from tilting experiments in the transmission electron microscope that the specimens had very irregular surfaces. Small tilts generally changed the caustic patterns visible at the edges. However, in spite of these complications, we conclude that useful information can be obtained by the use of this technique.

Various effects of irradiation on the properties of high temperature superconductors have been reported. We did not observe in twinned samples the orthorhombic to tetragonal transition reported by Eaglesham *et al.*⁸, although this has recently been described as a heating effect, rather than displacement damage¹¹. We suppose that the loss of superconductivity on electron irradiation and on ion thinning at room temperature reported here is likely to be the result of changing the oxygen stoichiometry of the $\text{YBa}_2\text{Cu}_3\text{O}_x$. However, our experiments also indicate that the loss of yttrium from the irradiated regions may also play an important part.

We acknowledge the assistance of Dr J. A. Wilson and Dr M. de Podesta in Bristol and the Institute of Physics in Beijing in obtaining the samples studied here. We wish to acknowledge the assistance provided in carrying out this work by several

members of the Electron Microscopy Group. Ms Yang Cui-Ying's research in Bristol is supported by the Academia Sinica.

Received 11 December 1987; accepted 27 January 1988.

1. Millman, W. S., Ullah, M., Han Zhilan, Maclaurin, M. & Tonner, B. *J. Mater. Res.* **2**, 543-546 (1987).
2. Marton, L. & Lachenbruch, S. H. *J. appl. Phys.* **20**, 1171-1182 (1949).
3. Berry, N. V. & Upstill, C. *Prog. Opt.* **18**, 257-346 (1980).
4. Blackman, M. & Grunbaum, E. *Proc. R. Soc. A* **241**, 508-521 (1957).
5. Pozzi, G. & Valdre, U. *Phil. Mag.* **23**, 745-762 (1971).
6. Ourmazd, A. *et al. Nature* **327**, 308-310 (1987).
7. Yang, C. Y., Zou, X. D., Zhon, Y. Q. & Fung, K. K. *Mod. Phys. Lett.* **1**, 67-71 (1987).
8. Eaglesham, D. J. *et al. Appl. Phys. Lett.* **51**, 457-459 (1987).
9. Capiluppi, C., Pozzi, G. & Valdre, U. *Phil. Mag.* **26**, 865-883 (1972).
10. de Wade, A. Th. A. M. *et al. preprint*, Eindhoven University of Technology.
11. Iijima, S., Ichihashi, T., Kubo, Y. & Tabuchi, J. *Jap. J. appl. Phys.* **26**, L1478-L1481 (1987).

A molecular sieve with eighteen-membered rings

Mark E. Davis*, Carlos Saldarriaga*†, Consuelo Montes*, Juan Garces‡ & Cyrus Crowder‡

* Department of Chemical Engineering, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061, USA

‡ The Dow Chemical Company, Midland, Michigan 48640, USA

In this letter we describe a novel family of aluminophosphate-based molecular sieves denoted as VPI-5. These molecular sieves are the first to contain rings consisting of greater than twelve tetrahedral atoms. The large pores of the VPI-5 materials are long channels circumscribed by rings containing eighteen tetrahedral atoms and possess free diameters of approximately 12-13 Å.

The first report of a zeolite molecular sieve was that of stilbite in 1756 (ref. 1). Since that time many natural and synthetic zeolites have been discovered. The term molecular sieve was coined by J. W. McBain² to describe porous materials which can act as sieves on a molecular scale, and is not limited to zeolites (that is, aluminosilicate materials). Recently, Union Carbide has synthesized aluminophosphate (AlPO_4) (ref. 3) and element-substituted aluminophosphate⁴ molecular sieves such as SAPO (silicon substitution) and MeAPO (metal substitution).

The aluminosilicate zeolites, some silica polymorphs⁵, and the aluminophosphate-based molecular sieves all consist of tetrahedrally coordinated atoms (T-atoms), such as Si, Al and P, which are linked to adjacent T-atoms by bridging oxygen atoms. The largest rings in natural zeolites¹ and synthetic molecular sieves^{1,4} contain 12 T-atoms. Thus, the free diameter available for adsorption is bounded by approximately 10 Å. The existence of the 12-membered ring has been known for quite some time. In fact, the natural zeolites gmelinite and faujasite both contain 12-membered rings¹.

Larger ring sizes are known. For example, the hydrated basic ferric oxyphosphate mineral caxenite⁶ contains rings of ~15 Å free diameter. However, caxenite possesses Fe(III)-oxygen octahedra. Also, hypothetical large ring structures have been predicted by Barrer and Villiger⁷, Smith and Dytrych⁸, and Meier⁹. Of importance to this work are the predictions of Smith and Dytrych who presented two three-dimensional nets with rings possessing from eight to an unlimited number of tetrahedra.

We report here the first molecular sieves which contain rings consisting of more than twelve T-atoms. Virginia Polytechnic Institute number 5 (VPI-5) is a family of aluminophosphate-based molecular sieves with the same three-dimensional

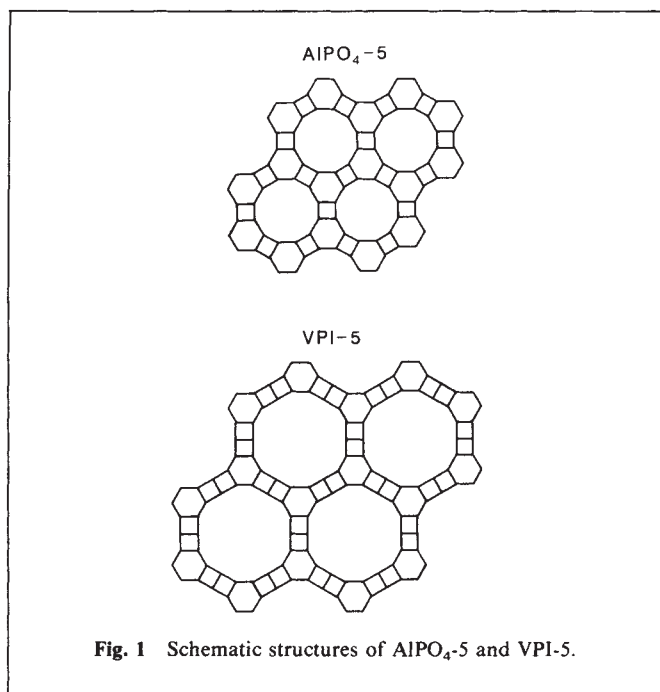


Fig. 1 Schematic structures of $\text{AlPO}_4\text{-5}$ and VPI-5.

topology. Figure 1 illustrates the topology of VPI-5. The structure of the silicon containing member (Si-VPI-5) was solved by X-ray powder-diffraction methods and the complete details of the refinement will be given elsewhere (C. Crowder *et al.* in preparation). In the refinement, the presence of silicon was ignored. The symmetry of this structure is hexagonal with unit cell dimensions of $a = 18.989$ Å, $c = 8.112$ Å and space group $P6_3$ cm ($P6_3$ /mmm). Notice that VPI-5 contains rings consisting of 18 T-atoms. The pores circumscribed by these rings have free diameter of ~12-13 Å.

Figure 1 illustrates the topology of $\text{AlPO}_4\text{-5}$ (ref. 10) and VPI-5. The 12 T-atom ring of $\text{AlPO}_4\text{-5}$ is comprised of six 4-membered rings and six 6-membered rings. If a 4-membered ring is inserted adjacent to each of the existing 4-membered rings in the $\text{AlPO}_4\text{-5}$ topology such that there are now pairs of 4-membered rings separating the 6-membered rings, then the VPI-5 topology is generated.

Smith and Dytrych⁸ label the $\text{AlPO}_4\text{-5}$ topology net 81. They also developed the concept of inserting the 4-membered rings into net 81 to give what they denote as net 81(1). Thus, the Smith and Dytrych net 81(1) is the topology of VPI-5. Net 81(1) was analysed by Meier⁹ and possesses a framework density of ~14. The framework density is defined to be the number of T-atoms per 1,000 Å (ref. 3). The true framework density of VPI-5 is 14.2. For comparison, the faujasite structure contains a framework density of 12.7 (ref. 9).

† Present address: Universidad de Antioquia, Medellin, Colombia.

The existence of large pores is confirmed experimentally by the adsorption of triisopropylbenzene and perfluorotributylamine which possess kinetic diameters of 8.5 and 10.5 Å, respectively. These molecules do not adsorb in known molecular sieves except for the case of triisopropylbenzene which adsorbs very slowly into faujasite type molecular sieves (equilibrium reached in approximately one day compared to approximately five minutes for VPI-5). VPI-5 can truly be classified as a molecular sieve because it can be reversibly dehydrated and re-hydrated at temperatures consistent with other known molecular sieves. Complete disclosure of the physicochemical properties of the VPI-5 materials is forthcoming (M. E. Davis *et al.*, in preparation). Thus, VPI-5 is the first molecular sieve with pores consisting of channels circumscribed by 18-member rings and having free diameters larger than 10 Å.

We thank the National Science Foundation and the Dow Chemical Company for support of this work through the Presidential Young Investigator Award to M.E.D.

Received 19 October; accepted 17 December 1987.

1. Breck, D. W. *Zeolite Molecular Sieves* (Wiley, New York, 1974).
2. McBain, J. W. *The Sorption of Gases and Vapors by Solids* (Rutledge, London, 1932).
3. Wilson, S. T., Lok, B. M., Messina, C. A., Cannan, T. R. & Flanigen, E. M. *J. Am. chem. Soc.* **104**, 1146–1147 (1982).
4. Flanigen, E. M., Lok, B. M., Patton, R. L. & Wilson, S. T. *New Developments in Zeolite Science and Technology* (eds Murakami, X., Iijima, A. & Ward, J. W.) 103–112 (Elsevier, Amsterdam, 1986).
5. Flanigen, E. M. *et al. Nature* **271**, 512–516 (1978).
6. Moore, P. B. & Shen, J. *Nature* **306**, 356–358 (1983).
7. Barrer, R. M. & Villiger, H. Z. *Kristallogr.* **128**, 352–370 (1963).
8. Smith, J. V. & Dytrych, W. J. *Nature* **309**, 607–608 (1984).
9. Meier, W. M. *New Developments in Zeolite Science and Technology* (eds Murakami, X., Iijima, A. & Ward, J. W.) 13–22 (Elsevier, Amsterdam, 1986).
10. Bennett, J. M., Cohen, J. P., Flanigen, E. M., Pluth, J. J. & Smith, J. V. *Am. chem. Soc. Symp. Ser.* **218**, 79–106 (1983).

Evidence from ^{29}Si NMR for the structure of mixed-layer illite/smectite clay minerals

Stephen P. Altaner, Charles A. Weiss Jr
& R. James Kirkpatrick

Department of Geology, University of Illinois, 1301 West Green St, Urbana, Illinois 61801, USA

Understanding the crystal structure of mixed-layer illite/smectite (I/S) is important to studies of petroleum migration¹ and maturation², geopressure development³, sandstone cementation⁴, and thermal histories of sedimentary⁵ and hydrothermally altered⁶ rocks. The crystallographic nature of I/S minerals is presently described by two competing theories, one based on Markov theory⁷ and the other based on transmission electron microscopy of thin crystallites, called 'fundamental particles'⁸. Here we present ^{29}Si NMR data obtained using an inversion/recovery technique⁹, which allows direct analysis of the Si/Al(4) ratios of individual illite and smectite components in I/S. The NMR results show that illite-rich I/S contains both high-charge and Al-rich (illite) and low-charge and Al-poor (smectite) structural environments. These results are best represented by the Markovian model of I/S. The NMR data are most consistent with the interpretation that the top and bottom tetrahedral sheets of both MacEwan crystallites and fundamental particles are smectitic. Fundamental particles are best interpreted as large crystallites of I/S that have been separated by osmotic swelling of Li^+ and Na^+ in smectite interlayers. Thermodynamic studies¹⁰ should consider illite-rich I/S as two phases, rather than as one phase, as suggested in previous investigations^{8,11–15}.

The Markovian model represents I/S as consisting of illite and smectite layers arranged in stacking sequences along the c^* -axis, with unit cell divisions centered on the interlayer (Fig.

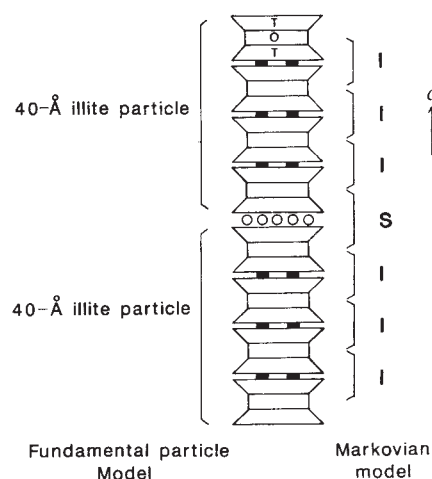


Fig. 1 Illite/smectite shown as a MacEwan crystallite (right) and an accumulation of two 40-Å-thick fundamental particles (left). T, tetrahedral sheet; O, octahedral sheet; black balls, interlayer K; white balls, interlayer water plus exchangeable cations. The fundamental particle theory predicts that the mineral shown in this figure contains one thermodynamic phase, illite, whereas the XRD approach based on Markov theory predicts the presence of two phases, illite and smectite. XRD modelling cannot determine whether the tetrahedral sheets at the top and bottom of the MacEwan crystallite are expandable (smectite) or non-expandable (illite). The crystallographic c^* or z -axis is perpendicular to the basal cleavage plane and is included here for reference; in monoclinic and triclinic phases, c^* is not parallel to c .

1). Illite contains dehydrated, K-rich interlayers with Al-rich tetrahedral sheets, whereas smectite contains hydrated interlayers with exchangeable cations and Si-rich tetrahedral sheets. Each stacking sequence is called a MacEwan crystallite. Reynolds and Hower¹⁶ applied Markov theory to calculate X-ray diffraction (XRD) patterns of I/S minerals. MacEwan crystallites probably terminate on the basal plane of oxygens of the tetrahedral sheets of the silicate layers rather than on one-half of an octahedral sheet¹⁷, but the composition of the tetrahedral sheet on the top and bottom of the MacEwan crystallite cannot be determined by XRD⁷.

Nadeau and co-workers^{8,11,13,14,18–20} have used transmission electron microscope (TEM) observations of platinum-shadowed samples to measure c^* -axis thicknesses of Li^+ - or Na^+ -exchanged I/S. They interpret their measurements as indicating that I/S consists of 'fundamental particles', whose c^* -axis dimensions are approximately integral multiples of 10 Å, and up to an order of magnitude smaller than the thicknesses of MacEwan crystallites inferred from XRD. A 10-Å-thick particle is interpreted as an isolated 2:1 layer and termed a fundamental smectite particle. Particles ≥ 20 Å in thickness are interpreted to be two or more 2:1 (tetrahedral sheet: octahedral sheet) layers linked by K atoms and are termed fundamental illite particles. Nadeau and co-workers suggest that accumulations of fundamental particles behave as MacEwan crystallites, such that during XRD analysis, interlayers within fundamental particles are perceived as illite layers and hydrated interfaces between particles are perceived as smectite layers (Fig. 1).

The most significant difference between these two models is that the fundamental particle theory, which has aroused considerable recent controversy^{21–23}, proposes that illite-rich I/S that contains fundamental particles ≥ 20 Å thick consists of only one thermodynamic phase, the mineral illite^{8,11–15}. This interpretation implies that an I/S sample has a constant composition within the 2:1 layers. In contrast, the approach based on Markov theory assumes that illitic I/S contains both expandable layers (smectite) and dehydrated layers (illite), which behave differently because they have different compositions⁷.

Table 1 Percentage of illite sites in I/S (P_I) and tetrahedral site compositions ($\text{Si}_{4-x}\text{Al}_x$) determined by NMR, XRD and TEM methods for sample RM30 and the rectorite sample of ref. 9.

	RM30*	Rectorite†
P_I (NMR)	78 ± 4	50 ± 4 (estimated)
P_I (XRD)	94 ± 2	50 ± 2
P_I (XRD _{adj})	86 ± 3	Not calculated
P_I (TEM)	91 ± 2	Not measured
$\text{Si}_{4-x}\text{Al}_x$ (XRF)	0.73 ± .02	0.74 ± .02
$\text{Si}_{4-x}\text{Al}_x$ (NMR)	0.71 ± .03	0.70 ± .03
$\text{Si}_{4-x}\text{Al}_x$ (XRD)‡	0.78 ± .03	Not calculated
$\text{Si}_{4-x}\text{Al}_x$ (XRD _{adj})‡	0.74 ± .03	Not calculated
$\text{Si}_{4-x}\text{Al}_x$ (TEM)‡	0.77 ± .03	Not calculated

Reported uncertainties in analyses are estimates of the root-mean-square precision.

* NMR and XRD data from this study; XRF data from ref. 14.

† Determined from data for sample B in ref. 9.

‡ Determined from $\text{Si}_{4-x}\text{Al}_x$ composition of illite and smectite sites from NMR data and % illite in I/S from XRD or TEM techniques.

^{29}Si magic-angle sample-spinning (MAS) NMR spectroscopy allows the accurate determination of the Si/Al(4) ratio of tetrahedral sites in aluminosilicates, including phyllosilicates²⁴⁻²⁸. For phyllosilicates with exchangeable interlayer sites, this method in conjunction with an inversion/recovery pulse sequence allows separate detection of Si in illite and smectite sites in 1% Mn-exchanged I/S⁹. Interlayer manganese (and also interlayer water³⁷, our data not shown) causes a more rapid relaxation of the Si nuclei in smectite, which are located adjacent to exchangeable interlayer sites. Here we analyse the Si/Al(4) composition of an I/S mineral with a very low iron content (sample RM30 of ref. 14) using these techniques. No other sample we have examined has a low enough iron content to allow similar analysis. We also present ordinary ^{29}Si MAS NMR spectra of several other samples which support our conclusions.

Comparison of the ^{29}Si MAS NMR spectrum of RM30 (Fig. 2a) with NMR spectra of pure illite and smectite^{26,28,29} reveals that RM30 contains peaks from Si sites in both smectite and illite. The peaks at -86.5 and -82.5 p.p.m. are primarily due to illite Q³(1Al) and Q³(2Al) sites, respectively (nomenclature defined in refs 24 and 25). The intensity at more negative chemical shifts is due to illite Q³(0Al) and smectite Q³(0Al) and Q³(1Al) sites. The observed XRD pattern of this sample is best modelled by a calculated pattern assuming 94% illite layers and R3 order (Fig. 3).

In the inversion/recovery ^{29}Si NMR spectra (Fig. 2b, c), signals from illite and smectite sites are completely separated from each other. All peaks are at different chemical shifts and there is no residual intensity for smectite in the illite spectrum or vice versa. Based on the binomial theorem and assuming Al avoidance³⁰, the calculated composition of the illite tetrahedral sites is $\text{Si}_{3.19}\text{Al}_{0.81}$ and that of the smectite tetrahedral sites is $\text{Si}_{3.66}\text{Al}_{0.34}$. Simulation of the ^{29}Si MAS NMR spectrum (Fig. 2a) using the method described in the Fig. 2 legend yields 78% illite and 22% smectite tetrahedral (Si + Al) sites. Based on these estimates of the tetrahedral compositions of the individual illite and smectite sites and the percentages of total illite and smectite in the sample, the calculated tetrahedral composition for the whole sample is $\text{Si}_{3.29}\text{Al}_{0.71}$ (Table 1). This composition is remarkably similar to that determined by X-ray fluorescence (XRF) ($\text{Si}_{3.27}\text{Al}_{0.73}$, Table 1) and strongly supports our interpretation of the NMR data. The estimated percentage of illite layers based on Markov theory, P_I (XRD) is 94% (Table 1), significantly greater than that determined by NMR, P_I (NMR). In addition, the tetrahedral Al substitution for 94% illite, $\text{Si}_{3.22}\text{Al}_{0.78}$, is significantly larger than the measured XRF value. These differences suggest that the composition of the top and

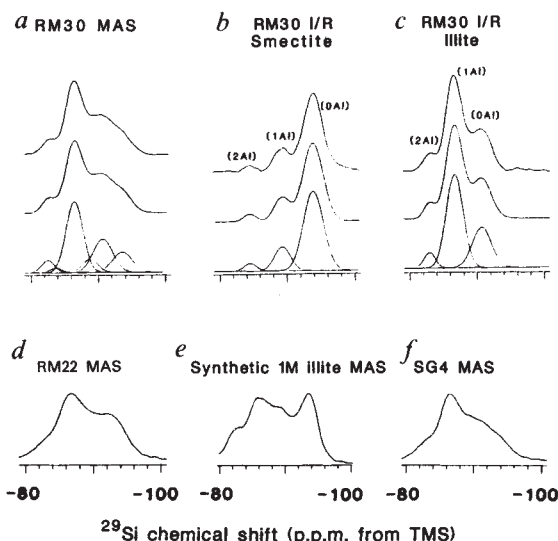


Fig. 2 ^{29}Si NMR spectra of illite-rich I/S obtained at 8.45 T (71.5 MHz) using the methods of Smith *et al.*³⁶, except that tetramethylsilane (TMS) was the external chemical-shift standard. *a*, ^{29}Si MAS spectrum (top), computer simulation (middle), and components of simulation (bottom) of sample RM30¹⁴ (2,236 scans, 4- μs pulse length, 10-s recycle time, 50-Hz gaussian line broadening, 10-kHz spectral window). Simulation contains both smectite and illite peak positions, based on the relative intensities of the peaks in *b* and *c*. In this simulation the relative intensities of the three smectite peaks are the same as observed in the inversion/recovery spectrum; the same is true for the three illite peaks. The observed relative intensities of the illite and smectite peaks are allowed to vary to obtain the illite/smectite ratio. The percentage of illite in I/S, P_I (NMR), is calculated from the peak areas in the simulation of this spectrum and an adjustment for the higher Si content in smectite as compared to illite. This calculation yields 22% smectite and 78% illite tetrahedral sites. *b*, ^{29}Si inversion/recovery spectra (top), computer simulations (middle), and components of simulations (bottom) for sample RM30 (1,000 scans, 5- μs pulse length, 5-s recycle time, 50-Hz gaussian line broadening, inversion/recovery pulse sequence of $180^\circ - \tau - 90^\circ$ -acquire). For *b*, $\tau = 1$ s; for *c*, $\tau = 120$ ms. This sample was saturated with Ca^{2+} and Mn^{2+} by reacting 0.5–1.0 g of clay with ~50 ml of a 1-M (1% Mn^{2+} /99% Ca^{2+}) solution at 60 °C for at least 2 h, repeated three times with fresh solutions, rinsed with de-ionized water each time, and air-dried at 60 °C. *b* contains signal from rapidly relaxing smectite sites which yield intensities of 77% Q³(0Al), 19% Q³(1Al) and 4% Q³(2Al). *c* contains signal from slowly relaxing illite sites and yields 31% Q³(0Al), 62% Q³(1Al), and 7% Q³(2Al). *d*, ^{29}Si MAS spectrum of sample RM22¹⁴ (2,230 scans, 4- μs pulse length, 10-s recycle time, 50-Hz exponential line broadening, 10-kHz spectral window). The sample is modelled by XRD assuming 91% illite layers and eight-layer-thick crystallites. Note signal for smectite Q³(0Al) sites near -94 p.p.m. *e*, ^{29}Si MAS spectrum of synthetic 1M illite (7,100 scans, 4- μs pulse length, 10-s recycle time, 50-Hz exponential line broadening, 10-kHz spectral window). The sample is modelled by XRD assuming 95–98% illite layers and five-layer-thick crystallites. The sample was synthesized by reacting equal weights of distilled water and gel with a cation composition of $\text{K}_{0.85}[\text{Al}_2(\text{Si}_{3.15}\text{Al}_{0.85})]$ at 350 °C and 1 kbar for 90 days. Note signal for smectite Q³(0Al) sites near -94 p.p.m. *f*, ^{29}Si MAS spectrum of sample SG4¹⁴ (1,632 scans, 4- μs pulse length, 10-s recycle time, 50-Hz exponential line broadening, 10-kHz spectral window). The sample is modelled by XRD assuming ~100% illite layers and 16-layer-thick crystallites. Note signal for smectite Q³(0Al) sites near -94 p.p.m.

bottom tetrahedral sheets of MacEwan crystallites (which are not examined by the XRD modelling technique⁷) are smectitic. The 'adjusted' XRD estimate of the percentage of illite layers, P_I (XRD_{adj}), is calculated assuming that MacEwan crystallites contain smectitic top and bottom tetrahedral sheets with the following expression: P_I (XRD_{adj}) = P_I (XRD) $x(N-1)/N$,

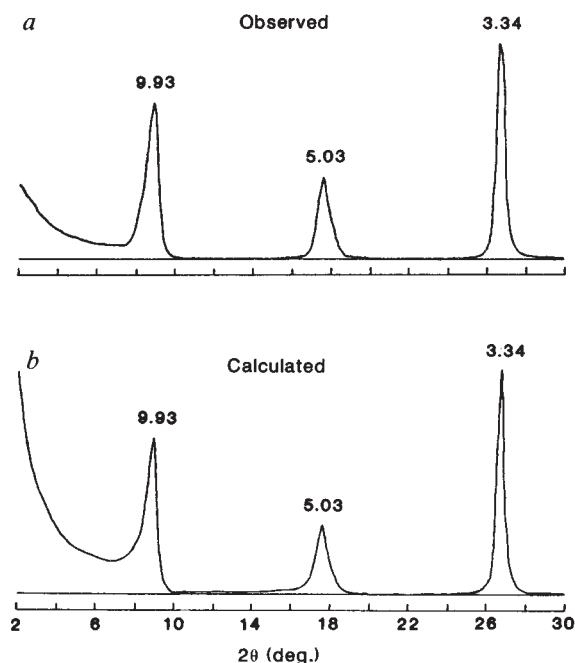


Fig. 3 Observed XRD pattern for ethylene-glycol-solvated sample RM30 (a) and calculated XRD pattern for I/S with 94% illite layers and perfect R3 order (b). Peaks are labelled with d -spacings in angstroms. To obtain the pattern in a, an I/S sample was oriented onto a glass slide by evaporation of a pipetted suspension and was saturated with ethylene glycol by equilibration in a glycol atmosphere at 60 °C for 12 h. XRD data were recorded with a Siemens D500 diffractometer, using Cu K α radiation, a step interval of 0.2° 2 θ and a counting time of 12 s. The calculated pattern (b) was generated using the NEWMOD computer program of R. C. Reynolds (Dartmouth College). Conditions assumed in the calculation were 94% illite layers in I/S, perfect R3 order, no iron, 0.8 K atoms, an average crystallite size of 12 layers, and maximum and minimum crystallite sizes of 36 layers and 1 layer, respectively, with the probability of occurrence of each crystallite size defined by $P(N) = \exp(-N/\delta)$, where $P(N)$ = probability of the occurrence of a crystallite with N 2:1 layers and δ = average crystallite size (total number of 2:1 layers).

where N = MacEwan crystallite thickness (total number of 2:1 layers). For RM30, $P_i(\text{XRD}_{\text{adj}})$ is 86% illite and the adjusted tetrahedral composition is $\text{Si}_{3.26}\text{Al}_{0.74}$, values similar to those derived from the NMR and XRF measurements. Weir and Rayner³⁸ made a similar adjustment to an XRD estimate of illite content for soil-derived I/S, which appeared to be much more smectitic based on chemical analysis.

TEM examination of sample RM30¹⁴ reveals no fundamental particles smaller than 37 Å; hence, fundamental particle theory predicts no smectite in the sample, a prediction clearly inconsistent with the NMR results. This discrepancy is reduced if the top and bottom tetrahedral sheets of each fundamental particle are assumed to be smectitic. The average thickness of particles measured by the TEM method for sample RM30 is 109 Å (ref. 14), and the assumption of smectitic tops and bottoms yields an estimated layer composition of 91% illite and a tetrahedral composition of $\text{Si}_{3.23}\text{Al}_{0.77}$. These values are closer to, but still significantly larger (in Al and illite content) than the NMR and XRF values. We therefore feel that there are likely to be smectite sites in the interiors of fundamental particles. For example, if the interlayer region between adjacent 2:1 layers is partly expandable and smectitic, but illitic and K-bonded in a large enough portion to resist physical separation, the illite content of this particle would be overestimated by the TEM method.

Additional support for our interpretation comes from the presence of the signal for Q³(0Al) smectite sites at -94 p.p.m. in the ²⁹Si MAS NMR spectra of three other I/S minerals with

illite contents ranging from 91% to nearly 100% as determined by comparison to calculated XRD patterns (Fig. 2d-f). These data support our suggestion that MacEwan crystallite size must be considered in the XRD estimate of smectite content. For example, the NMR spectrum of a synthetic illite (Fig. 2e) shows a minimum of 29% smectite sites (the intensity of the smectite Q³(0Al) peak at -94 p.p.m.), whereas XRD data reveal 2-5% smectite layers but an extremely small crystallite size (five layers). If the tops and bottoms of these small crystallites are smectitic, $P_i(\text{XRD}_{\text{adj}})$ is ~77% (~23% smectite layers), a value more compatible with the NMR data. Unfortunately, owing to the presence of iron in samples RM22 and SG4, and the small size of the synthetic illite sample, the inversion/recovery pulse sequence does not yield satisfactory data for them. Iron causes rapid relaxation of all of the Si sites and a small sample volume gives rise to a low signal to noise ratio. RM30 contains nearly no iron, whereas RM22 and SG4 have Fe₂O₃ contents of 0.89 and 0.42 wt %, respectively¹⁴.

The ²⁹Si inversion/recovery NMR data for a rectorite⁹ (short-range-ordered I/S with 50% illite layers) also supports our conclusions. The tetrahedral composition for this sample is $\text{Si}_{2.88}\text{Al}_{1.12}$ for the illite sites and $\text{Si}_{3.72}\text{Al}_{0.28}$ for the smectite sites (calculated from the NMR spectra in ref. 9). The tetrahedral charge of -1.12 developed on the illite sites represents solid solution towards margarite and is balanced by Na⁺ and Ca²⁺. Assuming 50% illite sites for this sample yields a total tetrahedral composition of $\text{Si}_{3.30}\text{Al}_{0.70}$, in excellent agreement with the composition determined by XRF ($\text{Si}_{3.26}\text{Al}_{0.74}$, Table 1). Mineralogically, this sample is nearly identical to sample Bre¹⁸, which contains almost exclusively 20-Å-thick fundamental particles. The fundamental particle model thus predicts a single phase of illite composition, whereas the NMR spectra show a large proportion of smectite sites.

The results of our study, therefore, support the Markovian description of I/S minerals if the top and bottom tetrahedral sheets of I/S crystallites are assumed to be smectitic. Our data, however, require significant modifications to the theory of fundamental particles. Contrary to predictions by the fundamental particle theory, illite-rich I/S minerals that yield fundamental particles at least 20 Å thick contain a significant proportion of low-charged and Si-rich smectite sites. Thus, the fundamental particle theory must be modified to include these sites.

We interpret the observation of thin fundamental particles^{8,11,13-15,18-20} to be due to cleavage of MacEwan crystallites along smectite interlayers, probably due to osmotic swelling of Na⁺ and Li⁺ in smectite interlayers during dispersion in solution. Previous studies have documented that Li- and Na-exchanged smectite in dilute solutions will completely dissociate into 10-Å 2:1 silicate layers^{31,32}.

The smectitic composition required for the top and bottom tetrahedral sheets of MacEwan crystallites may indicate that natural I/S consists of much thicker crystallites than the thicknesses inferred from XRD modelling. High-resolution lattice-fringe images of smectite³³ and I/S minerals³⁴ from bentonites reveal that these minerals consist of extremely thick crystallites (10 to >50 layers) that contain a mosaic of subparallel domains. Similar mosaic structure has been observed in TEM and XRD characterization of carbon minerals³⁵. The smectite interlayer regions, which are not observed by XRD, are candidates for planar defects in illitic I/S.

This research was supported by NSF and Texaco. We thank D. D. Eberl for samples used in this study.

Received 11 May; accepted 14 October 1987.

1. Powers, M. C. *Bull. Am. Ass. Petrol. Geol.* **51**, 1240-1254 (1967).
2. Johns, W. D. *Proc. int. Clay Conf.* Vol. 7 (eds van Olphen, H. & Veniale, F.) 655-664 (Elsevier, New York, 1982).
3. Bethke, C. M. *J. geophys. Res.* **91**, 6535-6545 (1986).
4. Boles, J. R. & Franks, S. G. *J. sedim. Petrol.* **49**, 55-70 (1979).
5. Hoffman, J. & Hower, J. *Spec. Publ. Soc. econ. Paleont. Miner.* **26**, 55-79 (1979).
6. Horton, D. G. *Contr. Miner. Petrol.* **91**, 171-179 (1985).

7. Reynolds, R. C. in *Crystal Structures of Clay Minerals and their X-ray Identification* (eds Brindley, G. W. & Brown, G.) 249–303 (Mineralogical Society, London, 1980).
8. Nadeau, P. H., Wilson, M. J., McHardy, W. J. & Tait, J. M. *Science* **225**, 923–925 (1984).
9. Barron, P. F., Slade, P. & Frost, R. L. *J. phys. Chem.* **89**, 3880–3885 (1985).
10. Garrels, R. M. *Clays Clay Miner.* **32**, 161–166 (1984).
11. Nadeau, P. H., Wilson, M. J., McHardy, W. J. & Tait, J. M. *Clay Miner.* **19**, 757–769 (1984).
12. Nadeau, P. H. & Bain, D. C. *Clays Clay Miner.* **34**, 455–464 (1986).
13. Wilson, M. J. & Nadeau, P. H. in *The Chemistry of Weathering* (ed. Drever, J. I.) 97–118 (Macaulay Institute for Soil Research, Aberdeen, 1985).
14. Eberl, D. D., Srodon, J., Lee, M., Nadeau, P. H. & Northrop, H. R. *Am. Miner.* **72**, 914–934 (1987).
15. Inoue, A., Kohyama, N., Kitagawa, R. & Watanabe, T. *Clays Clay Miner.* **35**, 111–120 (1987).
16. Reynolds, R. C. & Hower, J. *Clays Clay Miner.* **18**, 25–36 (1970).
17. Tettnerhorst, R. & Roberson, H. E. *Am. Miner.* **58**, 73–80 (1973).
18. Nadeau, P. H. *Clay Miner.* **20**, 499–514 (1985).
19. Nadeau, P. H., Tait, J. M., McHardy, W. J. & Wilson, M. J. *Clay Miner.* **19**, 67–76 (1984).
20. Nadeau, P. H., Wilson, M. J., McHardy, W. J. & Tait, J. M. *Mineralog. Mag.* **49**, 393–400 (1985).
21. Mackinnon, I. D. R. *Clays Clay Miner.* **35**, 74–76 (1987).
22. Ahn, J. H. & Peacor, D. R. *Clays Clay Miner.* **34**, 180–186 (1986).
23. Nadeau, P. H., Wilson, M. J., McHardy, W. J. & Tait, J. M. *Clays Clay Miner.* **35**, 77–79 (1987).
24. Kirkpatrick, R. J., Smith, K. A., Schramm, S., Turner, G. & Yang, W.-H. *Rev. Earth planet. Sci.* **13**, 29–47 (1985).
25. Lippmaa, E., Magi, M., Samoson, A., Engelhardt, G. & Grimmer, A.-R. *J. Am. chem. Soc.* **102**, 4889–4893 (1980).
26. Sanz, J. & Serratos, J. M. *J. Am. chem. Soc.* **106**, 4790–4793 (1984).
27. Lipsicas, M. *et al. Nature* **309**, 604–607 (1984).
28. Kinsey, R. A., Kirkpatrick, R. J., Hower, J., Smith, K. A. & Oldfield, E. *Am. Miner.* **70**, 537–548 (1985).
29. Weiss, C. A., Altaner, S. P. & Kirkpatrick, R. J. *Am. Miner.* **72**, 935–942 (1987).
30. Loewenstein, W. *Am. Miner.* **39**, 92–96 (1954).
31. Norrish, K. *Disc. Faraday Soc.* **18**, 120–134 (1954).
32. Foster, W. R., Savins, J. G. & Waite, J. M. *Clays Clay Miner.* **3**, 296–316 (1955).
33. Lee, J. H. & Peacor, D. R. *Clays Clay Miner.* **34**, 69–73 (1986).
34. Klimentidis, R. E. & Mackinnon, I. D. R. *Clays Clay Miner.* **34**, 155–164 (1986).
35. Ergun, S. *Phys. Rev. B1*, 3371–3380 (1970).
36. Smith, K. A., Kirkpatrick, R. J., Oldfield, E. & Henderson, D. M. *Am. Miner.* **68**, 1206–1215 (1983).
37. Barron, P. F., Slade, P. & Frost, R. L. *J. phys. Chem.* **89**, 3305–3310 (1985).
38. Weir, A. H. & Rayner, J. H. *Clays Clay Miner.* **10**, 173–187 (1974).

Long-term polarity stability and strength of the geomagnetic dipole

Poorna C. Pal* & Paul H. Roberts†

* Instituto Astronômico e Geofísico, Universidade de São Paulo, 01051 São Paulo, Brazil

† Institute of Geophysics and Planetary Physics, University of California, Los Angeles, California 90024, USA

Palaeomagnetically observed reversals so far provide the strongest evidence¹ that reversals originate from processes of core magnetohydrodynamics (MHD), thus furnishing a crucial constraint for geomagnetic dynamo modelling. The long-term variations in reversal frequency, favouring² independent mechanisms for the dynamo and its reversals, provide an added constraint. Here we present evidence for a correlation between polarity superchrons and palaeointensity data, and also marine-magnetic and MAGSAT anomalies. Our results indicate that the dipole moment is correlated to the reversal frequency, being appreciably greater during the fixed polarity superchrons. This new information derived from the reversal record provides fresh insight into the workings of the geodynamo, and also explains why the oceanic crust associated with the Cretaceous quiet zone carries such a strong thermoremanence.

Even though the occurrence of reversals is by now a well-established fact, largely because of the interpretation³ of marine-magnetic data, their theoretical modelling has remained difficult despite intense efforts^{4–13}. So far, the most successful geodynamo models have been kinematic, that is, ones in which non-decaying solutions of the magnetic induction equation are obtained for assumed velocity fields, v ; only highly truncated subsets of the full MHD dynamo equations have as yet been integrated. Although the magnetic field, B , in some of these models does possess oscillatory régimes, simulating the irregularly occurring geomagnetic reversals, the duration τ_p of the polarity epochs in these solutions is intermediate between the electrodynamic free

decay time, τ_σ , and the dynamic timescale, τ_d . For the core, τ_σ may be $\sim 5 \times 10^{3-4}$ years, and a dynamic timescale that can make τ_p long enough to mimic the observed reversal record is difficult to identify. Enhanced palaeogeophysical constraints on modelling are therefore readily desirable, as a recent examination² of long-term reversal behaviour readily demonstrates. Because of the inferred^{14,15} collapse of the dipole field during the reversal process, it is interesting to examine whether the palaeointensity record also contains any comparable characteristics that can aid these modelling efforts. We examine here this possibility, to which the interpretation^{16–18} of intermediate-wavelength satellite magnetic anomalies in terms of significant changes in the thermoremanence of the oceanic crust lends added credence.

Concerning the long-term changes in reversal frequency, the existing record reveals intermittent reversals throughout the entire geological past¹, except for 35–70-Myr-long fixed polarity superchrons^{19,20}, such as those that occurred during the Cretaceous (N) (83–118 Myr), Permian (R) (250–320 Myr) and perhaps also the Cambrian (R) (535–590 Myr), although the magnetostratigraphically calibrated record barely covers the past 165-Myr period. A dominating aspect of this chronology, apparent in Fig. 1a where these reversals are seen through a 15-Myr window, is a progressive diminution^{21–23} in reversal frequency as the Cretaceous N polarity superchron developed.

The existing palaeointensity estimates for the past 180-Myr period, summarized in Fig. 1b, display an interesting negative correlation with this reversal behaviour. The palaeointensity coverage of the Cretaceous N polarity superchron is rather scant; only the data from Pacific seamounts²⁴, Damodar valley dikes²⁵, the Mount Dromedary complex²⁶, Rajmahal lavas²⁵ and Armenian porphyrites²⁷ being available. And yet, whereas the estimates from the first two of these are comparable with those for the late (0–25 Myr) and early (50–65 Myr) Cenozoic, but appreciably higher than those for the middle Cenozoic (25–50 Myr) and the upper-middle Mesozoic (130–190 Myr), those from the last two of these formations are significantly greater than all the other estimates considered here. The Mount Dromedary complex²⁶ (93–98 Myr), having mixed directions, gives lower intensities than all palaeointensity data included in Fig. 1b. But this finding can be questioned because of the detection of a partially self-reversing thermoremanence²⁸ in some of the samples from this formation; it has therefore been excluded. The remaining data, summarized in Fig. 1b, indicate an appreciably stronger geomagnetic field during the Cretaceous N polarity superchron than during the adjacent superchrons of frequent reversals.

The palaeointensity estimates available from the Permian R polarity superchron also support this inference. For instance, data²⁹ from the highly oxidized Exeter and Esterel volcanics, both belonging to this superchron, indicate intensities that are 30–50% greater than the mean palaeointensity estimates for the remainder of the Devonian–Triassic interval³⁰, as are those from the R polarity Bohemian³¹ ignimbrites (Permian) and cassiterites (Permo–Carboniferous). But the most interesting evidence here comes from the palaeointensity estimates based on the resedimentation of Russian red clays³¹. As shown in Fig. 1c, these intensity estimates, when taken in conjunction with the results summarized in Fig. 1b, indicate conspicuous increases in dipole moment during fixed polarity superchrons.

It should be recognized that the palaeointensity record suffers from two chief limitations. One is that of large fluctuations, ≤ 40 –50% of the mean value, which occur within any geological interval. For instance, during the past 12,000 years the geomagnetic dipole moment has fluctuated³² around 5 – 12×10^{22} A m², with mean (x) and standard deviation (s.d.) of 8.75×10^{22} A m² and 1.58×10^{22} A m². Similar estimates of x and s.d. from the past 5 Myr period³³ are 8.67×10^{22} A m² and 3.63×10^{22} A m², respectively, which are comparable with the corresponding values of 8.9×10^{22} A m² and 3.4×10^{22} A m² obtained from an

Fig. 1 *a*, Frequency of geomagnetic reversals during the past 165 Myr when the magnetostratigraphically calibrated chronology¹⁹ is seen through a 15-Myr-wide rectangular window. Although reversal spurts²³ do exist in this record, spaced about 30 Myr apart, the dominant trend here is one of progressive diminution in reversal frequency towards the Cretaceous N polarity superchron. *b*, Estimates of palaeogeomagnetic dipole moment for the past 180 Myr, relative to the present (vertical bars denote ranges of values, horizontal lines age ranges, and solid circles lognormal means), compiled from the following data (note the increase in dipole moment in the middle Cretaceous, corresponding to the N polarity superchron): (1) Past 5 Myr³³; (2) Alturitas lavas⁵⁷; (3) Icelandic lavas⁵⁸⁻⁶⁰; (4) Komoro igneous suite⁶¹; (5) Transcarpathian volcanics²⁷; (6) Steens Mountain lavas⁶²; (7) Bohemian massif³¹; (8) San Juan volcanics⁶³; (9) German Tertiary igneous suite⁶⁴; (10) Scottish dikes⁶⁴; (11) Antrim igneous suite²⁵; (12) Deccan lavas^{25,64-66}; (13) Damodar valley dikes²⁵; (14) Pacific seamounts²⁴; (15) Armenian porphyrites²⁷; (16) Rajmahal lavas²⁵; (17) Canadian gabbros⁶⁷; (18) Ferrar dolerites²⁶; (19) Red Hill granophyre²⁶; (20) Freetown igneous complex⁶⁸; (21) Stromberg lavas⁶⁴. *c*, Palaeointensity estimates based on the resedimentation of Russian red clays³¹ (solid circles), relative to the present, and the mean data from *b* (open circles) clearly suggest that the geomagnetic field was stronger during the Cretaceous N and Permian R polarity superchrons than during the superchrons of frequent reversals. Note the excellent match between the Russian data and the global estimates for the Cenozoic, the only interval when they overlap.

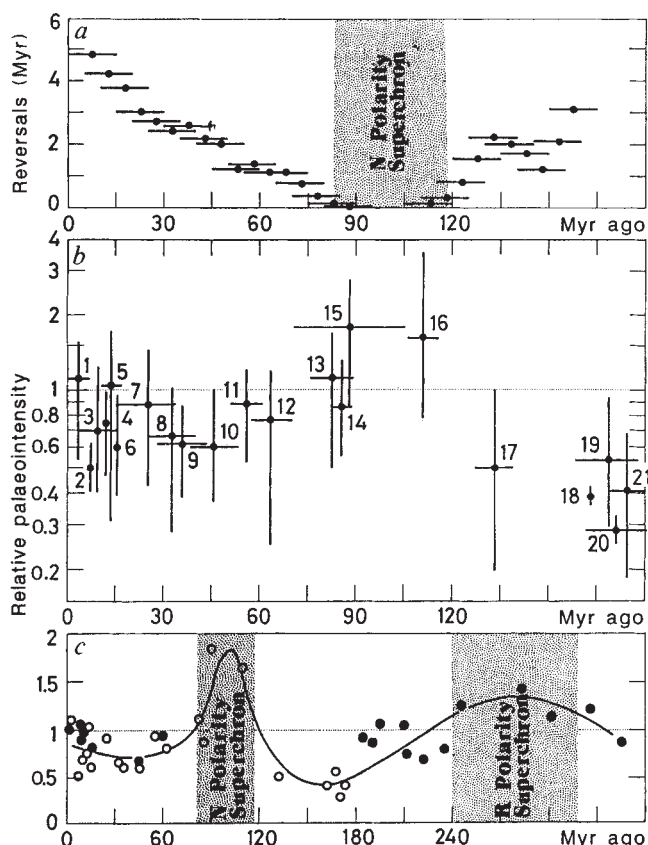
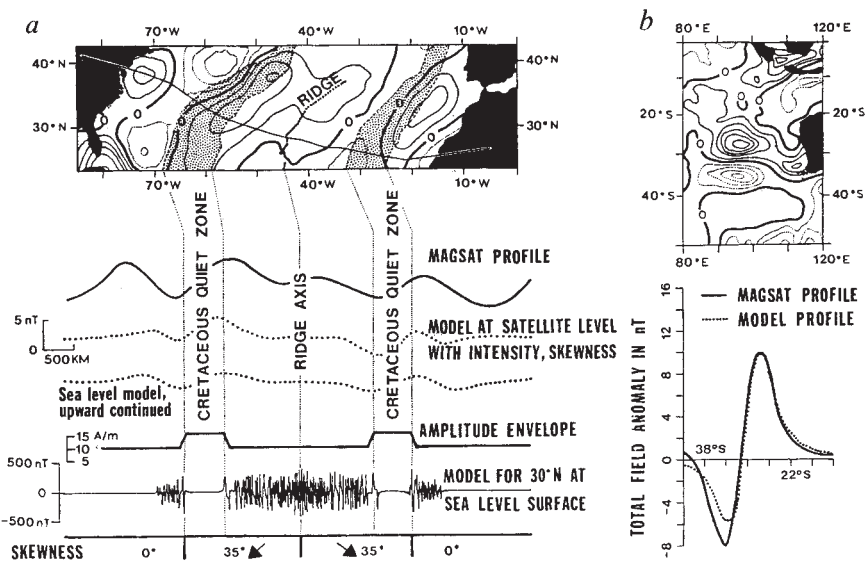


Fig. 2 *a*, The MAGSAT field over the central-north Atlantic (top) clearly shows a strong gradient, up to 1 nT deg^{-1} , over the middle Cretaceous oceanic crust. Solid contours are positive, broken ones negative; the contour interval is 2 nT . The modelling of the MAGSAT profile (bottom) favours a 50% stronger magnetization of the ocean floor associated with the Cretaceous quiet zone, and a 35° anomalous skewness for this and the younger ocean floor. Redrawn from LaBrecque and Raymond¹⁸. *b*, The MAGSAT field over the Broken Ridge, Indian Ocean, clearly shows (top) a strong dipole anomaly, having a gradient of about 2 nT deg^{-1} , over this sub-oceanic feature magnetized during the Cretaceous N polarity superchron (contour notations are the same as in *a*). Also shown here (bottom) is the modelling of a MAGSAT profile over this ridge, along 96°E , assuming a source inclination parallel to the present field. But the magnetization required by the model is almost three times the usual magnetization of oceanic crust, and favours an assumption of strong thermoremanence. (The anomalous skewness⁴¹ of $\sim 30^\circ$ can then explain the discrepancy between the expected and model inclinations.) Redrawn from Johnson⁴⁰.



earlier³⁶ analysis of the past 10 Myr (ref. 35). The second problem³⁶ consists of difficulties in replication, and in the mineralogical changes caused by laboratory-induced oxidation. Because of these, the palaeointensity data are most commonly used only as relative measures. Thus, although Fig. 1*b*, *c* indicate that the field strength increases with its protracted stability during a fixed polarity superchron, this suggestive trend requires independent corroboration, before its implications are seriously investigated.

Interestingly, such an independent corroboration is actually available from the marine and satellite magnetic anomaly profiles over the Cretaceous oceanic crust. The Cretaceous N polarity superchron occurs as a quiet zone of high amplitude, terminating in anomaly 34 of the marine-magnetic profiles. It

can be modelled only by assuming significant palaeointensity modulations. For example, the modelling of anomalies M-22 (155.77–156.71 Myr)¹⁹ to 34 in the Atlantic³⁷⁻³⁹ requires the underlying oceanic crust to have almost twice the remanence of the adjoining oceanic crust. The corresponding MAGSAT anomalies over the Cretaceous quiet zone, particularly over the Atlantic ocean¹⁸ (Fig. 2*a*) and Indian ocean (Broken Ridge)⁴⁰ (Fig. 2*b*), demonstrate this phenomenon best: they are too strong to be modelled by induction and/or remanent magnetization of a normal oceanic crust alone. The viscous effects resulting from the prolonged exposure to a fixed polarity during the Cretaceous N polarity superchron could have contributed to this strong remanence^{40,41}, but the inference that strong thermoremanence

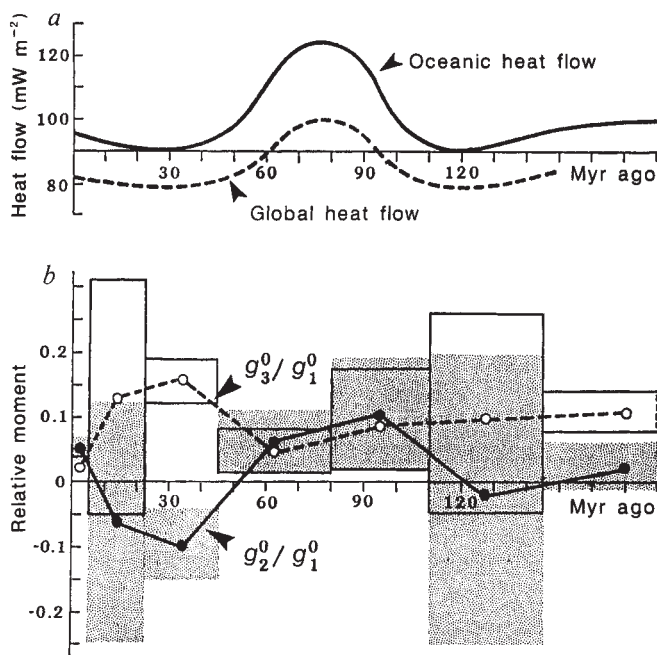


Fig. 3 *a*, The variation of oceanic and global geothermal fluxes since mid-Jurassic times⁵³. A markedly high heat flow occurred from 110 to 40 Myr ago, a period comprising the Cretaceous N polarity superchron and the first half of the present superchron of frequent reversals. This record exhibits no clear correlation with the palaeogeomagnetic records of panels *a* and *b* in Fig. 1. *b*, The variations of quadrupole (g_2^0) and octupole (g_3^0) components of the geomagnetic field, relative to the dipole moment (g_1^0), for the past 180 Myr, as reported by Lee and Lilley⁴⁶ based on Lee's spherical harmonic analysis of the global palaeomagnetic palaeopole data. Note the apparent association of a reversal of hemispheric symmetry with the period of increased heat flow shown in *a*.

was acquired in a relatively strong magnetizing field would be more consistent with the palaeomagnetic palaeointensity estimates that we have described above.

We therefore infer from the palaeomagnetic record that the geomagnetic dipole moment is linked to its protracted polarity stability, increasing appreciably during a fixed polarity superchron. Why such a correlation should exist is a theoretical problem of core MHD, involving questions about the magnitude of the poloidal field, that are not easily addressed. Clearly they cannot be fully answered by the equations of kinematic dynamo theory, which are linear in the field, but kinematic theory can say the following. Although the core has both poloidal (B_p) and toroidal (B_t) fields, only the former can be present at the Earth's surface. In the core, B_t is larger than B_p , by a factor of the order of the magnetic Reynolds number $R (=UL\mu\sigma)$, where U , L , μ and σ denote flow speed, length scale, magnetic permeability and electrical conductivity in the fluid core). In the so-called strong-field geodynamo models³⁶, R is large and the toroidal field dominates; in the weak-field models^{43,44}, $R \sim 1$ and the two fields are comparable in magnitude. Hide⁴⁵ suggested that, by analogy with the kinematics of perfect conductors ($R = \infty$), in which an external field cannot be created unless one was originally present, B_p should vanish as $R \rightarrow \infty$; he proposed that $B_p \sim R^{-1/2} B$, where B is a scale for the magnetic field strength, cannot be determined kinematically but is approximately independent of R . Because $B_t/B_p \approx R$, $B_t \sim R^{1/2} B$.

The field scale B is decided by the dynamics, which also determines R . If the geodynamo were working stably⁴⁶ in a weak-field régime, equilibration would require a field strength B such that the Hartmann number $M (=BL(\sigma/\rho\nu)^{1/2})$, where ρ and ν are the density and kinematic viscosity of the core) ~ 1 . But the very low estimate of B so obtained, only a few tens of

microtesla, indicates the geophysical relevance of strong-field régimes. For these, the equation governing the thermal wind indicates⁴⁷ that the Coriolis and Lorentz forces balance one another, so that $B_t^2/\mu L \approx 2\Omega\rho U$, that is, $B_t^2 \approx 2\Omega UL\mu\rho = 2\Omega\rho R/\sigma$. Because $B_t \sim R^{1/2} B$, the Elsasser number $\Lambda (= \sigma B^2/2\Omega\rho)$ based on B , will then be of order unity, suggesting that B is independent of U , as Hide⁴⁵ supposed. The results in Fig. 1*b, c* then suggest that R and U are probably smaller during a fixed-polarity superchron, when the dipole tends to be stronger than during a mixed-polarity superchron.

A reduced R may well signal enhanced convective activity in the core during a fixed-polarity superchron, for the following reason. Strong-field dynamos become regenerative at order 1 values of the dynamo number⁴⁸, $D = RR_\alpha$, where $R_\alpha = \alpha L\mu\sigma$ is the magnetic Reynolds number associated with the α -effect⁴⁰, indispensable to almost all dynamo models and attributed⁴⁹ to the inductive effects of waves that drift longitudinally round the core, waves that may be associated⁵⁰ with the observed westward drift of the geomagnetic field. A reduction in R at fixed D implies an increase in R_α , and a concomitant increase in the wave amplitude, as might well result from an increase in the energy input from convection⁵.

As a mechanism to explain the long-term variations in the palaeogeomagnetic record, one could appeal to an episodic thermal history of the core-mantle boundary, as has been recently suggested from an analysis² of the reversal record. Note the sensitive role that all forms of core-mantle coupling play in the 'model-Z' dynamo⁴², for instance. Besides being associated with stronger dipole fields, it is likely that relatively vigorous core motions during fixed-polarity superchrons would be associated with the thermal disintegration of parts of the D'' layer as mantle plumes, leading to an increase in the heat flux from the core to the mantle and to a corresponding surge in surface heat flux⁵². A comparison of geothermal and geomagnetic records should be revealing, although the large thermal inertia of the mantle might deny their explicit correlation. Note that, although the geothermal flux was appreciably high from about 110 to 40 Myr ago⁵³ (Fig. 3*a*), this correlates less directly with the observed reversal frequency and palaeointensity variations than with the results of a recent⁵⁴ spherical harmonic analysis of the palaeomagnetic palaeopoles (Fig. 3*b*) that suggests a reversal in the north-south hemispheric asymmetry during this interval. As the symmetry of heat flux relative to the equator may perhaps be reversed by a new thermal cataclysm, depending on the latitude on the core-mantle boundary at which it takes place, such a cataclysm might also reverse the hemispheric asymmetry of field, without greatly affecting the geomagnetic dipole moment or the polarity history. We note that two earlier studies^{55,56} have suggested weaker quadrupole moments during the late Cretaceous.

Our discussion of dynamic aspects of the geodynamo has, in its simplicity, omitted effects that might possibly be significant: for example, some dependence of B on the convective energy flux, which was disregarded in the estimate $B \approx (2\Omega\rho/\sigma)^{1/2}$. Further advances in theory are apparently needed before a realistic modelling of the observed palaeogeomagnetic record will become possible. What the above presentation establishes, however, is that the palaeointensity data, and not the reversal record alone, should be regarded as integral to such a record. Also, besides providing an interesting insight into the long-term behaviour of the geomagnetic field, the results we have presented probably provide the first instance where the interpretations of marine-magnetic and MAGSAT data have been used to establish the reliability of palaeomagnetic palaeointensity estimates.

P.C.P. thanks Professor Margaret Kivelson and the Department of Earth and Space Sciences in the University of California at Los Angeles, and also Instituto Astronômico e Geofísico, Universidade de São Paulo, for facilities. P.H.R. acknowledges the National Science Foundation. We also thank Professors Kenneth Creer, Frank Stacey, Ronald Merrill and Igor Pacca.

Received 29 July; accepted 26 December 1987.

1. Jacobs, J. A. *Nature* **323**, 296 (1986); *Reversals of the Earth's Magnetic Field* (Hilger, Bristol, 1984); *Q. J. R. astr. Soc.* **27**, 328–350 (1986).
2. McFadden, P. L. & Merrill, R. T. *Physics Earth planet. Inter.* **43**, 22–33 (1986).
3. Vine, F. J. & Matthews, D. H. *Nature* **199**, 947–949 (1963).
4. Crossley, D., Jensen, O. & Jacobs, J. A. *Physics Earth planet. Inter.* **42**, 143–153 (1986).
5. Robbins, K. A. *Math. Proc. Camb. phil. Soc.* **82**, 309–325 (1977).
6. Krause, F. & Roberts, P. H. *Adv. Space Res.* **1**, 231–240 (1980).
7. Rikitake, T. *Proc. Camb. phil. Soc.* **54**, 89–105 (1958).
8. Allan, D. W. *Proc. Camb. phil. Soc.* **58**, 671–693 (1962).
9. Cook, A. E. & Roberts, P. H. *Proc. Camb. phil. Soc.* **68**, 547–569 (1970).
10. Nozières, P. *Physics Earth planet. Inter.* **17**, 55–74 (1978).
11. Ito, K. *Earth planet. Sci. Lett.* **51**, 451–456 (1980).
12. Honkura, Y. & Shimizu, M. *J. Geomagn. Geoelectr., Kyoto* **37**, 1147–1155 (1980).
13. Kono, M. *Geophys. Res. Lett.* **14**, 21–24 (1987).
14. Prevot, M., Mankinen, E. A., Gromme, C. S. & Coe, R. S. *Nature* **316**, 230–234 (1985).
15. Valet, J. P., Lai, C. & Tucholka, P. *Nature* **322**, 27–31 (1986).
16. LaBrecque, J. L. & Cande, S. C. *J. geophys. Res.* **89**, 11,241–11,244 (1984).
17. LaBrecque, J. L., Cande, S. C. & Jarrard, R. D. *J. geophys. Res.* **90**, 2,549–2,564 (1985).
18. LaBrecque, J. L. & Raymond, C. A. *J. geophys. Res.* **90**, 2,565–2,575 (1985).
19. Harland, W. B., Cox, A., Llewellyn, P. G., Pickton, C. A. G., Smith, A. G. & Walters, R. A. *Geologic Time Scale* (Cambridge University Press, 1982).
20. Khrumov, A. N. & Rodionov, V. P. in *Global Reconstruction and the Geomagnetic Field During the Palaeozoic*, 99–115 (eds McElhinny, M. W., Khrumov, A. N., Ozima, M. & Valencia, D. A.) (Reidel, Dordrecht, 1981).
21. Lowrie, W. & Kent, D. V. *Earth planet. Sci. Lett.* **62**, 305–313 (1983).
22. McFadden, P. L. & Merrill, R. T. *J. geophys. Res.* **89**, 3354–3362 (1984).
23. Pal, P. C. & Creer, K. M. *Nature* **320**, 148–150 (1986).
24. Ozima, M., Ozima, M. & Kaneoka, I. *J. geophys. Res.* **73**, 711–723 (1968).
25. Radhakrishnamurty, C., Likhite, S. D. & Sahasrabudhe, P. W. *Physics Earth planet. Inter.* **13**, 289–300 (1977).
26. Briden, J. C. *Nature* **212**, 246–247 (1966).
27. Bol'shakov, A. S. & Solodovnikov, G. M. *Izv. Earth Phys. (Eng. trans. of Izv. Akad. Nauk SSSR)* **16**, 602–614 (1980).
28. Irving, E. *Palaeomagnetism* (Wiley, New York, 1964).
29. Senanayake, W. E. & McElhinny, M. W. *J. geophys. Res.* **85**, 85–96 (1983).
30. Smith, P. J. in *Palaeogeophysics*, 79–90 (ed. Runcorn, S. K.) (Academic, London, 1970).
31. Petrova, G. N. & Burlatskaya, S. P. in *Magnetic Field and the Processes in the Earth's Interior*, 1–200 (eds Bucha, V. et al.) (Academia, Prague, 1983).
32. McElhinny, M. W. & Senanayake, W. E. *J. geomagn. Geoelectr., Kyoto* **34**, 39–47 (1982).
33. McFadden, P. L. & McElhinny, M. W. *J. geomagn. Geoelectr., Kyoto* **34**, 163–179 (1982).
34. Kono, M. *Earth planet. Sci. Lett.* **5**, 140–150 (1971).
35. Walton, D. *Nature* **328**, 789–791 (1987).
36. Merrill, R. T. & McElhinny, M. W. *The Earth's Magnetic Field* (Academic, London, 1983).
37. Hayes, D. E. & Rabinowitz, P. D. *Earth planet. Sci. Lett.* **28**, 105–115 (1975).
38. Rabinowitz, P. D., Cande, S. C. & Hayes, D. E. *Init. Rep. DSDP* **43**, 879–885 (1979).
39. Vogt, P. R. & Einwich, A. M. *Init. Rep. DSDP* **43**, 857–876 (1979).
40. Johnson, B. D. *J. geophys. Res.* **90**, 2,640–2,646 (1985).
41. Cande, S. C. *Earth planet. Sci. Lett.* **40**, 275–286 (1978).
42. Braginskii, S. I. *Geomagn. Aeronomy* **15**, 149–156 (1975); *Physics Earth planet. Inter.* **11**, 191–199 (1976); *Geomagn. Aeronomy* **18**, 225–231 (1978).
43. Steenbeck, M. & Krause, F. *Astr. Nachr.* **291**, 271–286 (1969).
44. Busse, F. H. *Geophys. J. R. astr. Soc.* **42**, 437–459 (1975).
45. Hide, R. *Phil. Trans. R. Soc. A* **306**, 223–234 (1982).
46. Soward, A. M. *Physics Earth planet. Inter.* **20**, 134–151 (1979).
47. Taylor, J. B. *Proc. R. Soc. A* **274**, 274–283 (1963).
48. Parker, E. N. *Cosmical Magnetic Fields* (Clarendon, Oxford, 1979).
49. Braginskii, S. I. *Geophys. astrophys. Fluid Dyn.* **14**, 189–208 (1980); *Geomagn. Aeronomy* **4**, 698–712 (1964); **7**, 851–859 (1967).
50. Hide, R. *Phil. Trans. R. Soc. A* **259**, 615–647 (1966).
51. Roberts, P. H. & Soward, A. M. *A. Rev. Fluid Dyn.* **4**, 117–154 (1972).
52. Loper, D. E. & McCartney, K. *Geophys. Res. Lett.* **13**, 1,525–1,528 (1986).
53. Sprague, D. & Pollack, H. N. *Nature* **285**, 393–395 (1980).
54. Lee, S. & Lilley, F. M. *J. Geomagn. Geoelectr., Kyoto* **38**, 797–806 (1986).
55. Coupland, D. H. & Van der Voo, R. *J. geophys. Res.* **85**, 3,529–3,548 (1980).
56. Livermore, R. A., Vine, F. J. & Smith, A. G. *Geophys. J. R. astr. Soc.* **79**, 939–961 (1984).
57. Coe, R. S. & Gromme, C. S. *J. Geomagn. Geoelectr., Kyoto* **25**, 415–435 (1973).
58. Wilson, R. L., Dagley, P. & McCormack, A. G. *Geophys. J. R. astr. Soc.* **28**, 213–226 (1972).
59. Shaw, J., Dagley, P. & Musset, A. *Geophys. J. R. astr. Soc.* **68**, 211–225 (1982).
60. Kristjansson, L. & McDougall, I. *Geophys. J. R. astr. Soc.* **68**, 273–285 (1982).
61. Coe, R. S. *J. geophys. Res.* **72**, 3247–3262 (1967).
62. Prevot, M. & Watkins, N. D. *Cr. heb. Séanc. Acad. Sci., Paris D* **276**, 2431–2433 (1973).
63. Kono, M. & Uneno, N. *Physics Earth planet. Inter.* **13**, 305–314 (1977).
64. Smith, P. J. *Geophys. J. R. astr. Soc.* **12**, 321–362 (1967).
65. Kono, M. *J. geophys. Res.* **79**, 1135–1141 (1974).
66. Day, R., Dunn, J. R. & Fuller, M. *Physics Earth planet. Inter.* **13**, 301–304 (1977).
67. Schwartz, E. J. & Symons, D. T. A. *Physics Earth planet. Inter.* **1**, 122–128 (1968).
68. Briden, J. C., Henthorn, D. I. & Rex, D. C. *Earth planet. Sci. Lett.* **12**, 385–391 (1971).

Granulite facies Nd-isotopic homogenization in the Lewisian complex of northwest Scotland

M. J. Whitehouse*

Department of Earth Sciences, University of Oxford, Parks Road, Oxford OX1 3PR, UK

A published Sm–Nd whole-rock isochron of $2,920 \pm 50$ Myr, obtained from a wide range of lithologies in the Lewisian complex of north-west Scotland, was interpreted¹ as the time of protolith formation. This date is ~260 Myr older than estimates for the timing of high-grade metamorphism in the complex at ~2,660 Myr^{2,3}, and this period is considered to represent the duration of the Lewisian crustal accretion–differentiation super-event (CADS)⁴. Here we give new Sm–Nd data, obtained specifically from granulite facies tonalitic gneisses, that yield a date of $2,600 \pm 155$ Myr. Although depleted-mantle model ages (t_{DM}) suggest >200 Myr of pre-metamorphic crustal residence, the regression date and its associated initial Nd-isotopic parameters demonstrate Nd-isotopic homogenization during the high-grade event, as well as the probability of general rare-earth-element (REE) mobility. Models for selective element depletion in the complex have previously assumed REE immobility since 2,920 Myr, but the data presented here suggest that a reappraisal of the depletion mechanism is required.

The late-Archaeon (~2,900 Myr) Lewisian complex of mainland NW Scotland is divided into three main regions: the granulite facies central region and the amphibolite facies northern and southern regions. The present disposition of these regions is generally interpreted in terms of juxtaposed crustal levels⁵. Throughout the complex, the predominant lithology is tonalitic to granodioritic grey gneiss, within which occur lenses of mafic-ultramafic rocks and metasedimentary assemblages. Large-ion lithophile element (LILE) studies^{5,6} have shown that there is a progressive increase in the degree of depletion with

increasing grade. K/Rb ratios, which are higher than in other granulite facies terrains⁷, are a good indicator of the highly LILE-depleted nature of this region. It has recently been suggested⁸ that low LILE concentrations might, in part, reflect original source region characteristics; the prograde increase in LILE depletion probably represents superimposed geochemical differentiation.

The timing of Badcallian high-grade metamorphism in the complex is well established, with Scourie granulites yielding a Pb/Pb whole-rock isochron date of $2,680 \pm 60$ Myr³ and a U–Pb zircon date of $2,660 \pm 20$ Myr². Comparable dates from other parts of the complex (Rona⁹, Harris¹⁰ and the southern Outer Hebrides¹¹) illustrate the regional distribution of this event.

Hamilton *et al.*¹ obtained a combined Sm–Nd whole-rock regression date of $2,920 \pm 50$ Myr from tonalitic to granodioritic gneisses from both amphibolite and granulite facies regions, as well as mafic gneisses from the Drumbeg layered mafic-ultramafic complex. This date, with its initial $^{143}\text{Nd}/^{144}\text{Nd}$ ratio close to contemporaneous depleted-mantle reservoir (DM; ϵ_{Nd} (2,920) $\approx +2$), was interpreted as the time of differentiation of the igneous protoliths of the gneiss complex from the mantle. The ~260 Myr difference between this date and the time of regional metamorphism was considered to represent the duration of the Lewisian CADS. The assumption that the $2,920 \pm 50$ Myr date represents protolith formation, together with general observations of REE immobility during high-grade metamorphism^{12–14}, has led several workers to propose REE immobility during the Badcallian event^{6,15,16}. In consequence, mechanisms for selective element depletion in which REE could be mobilized (for instance CO_2 -rich fluid¹⁷; melt phase¹⁸) have previously been ruled out. The combination of samples with such clearly different histories has been criticized on the grounds that it could result in a biased age estimate¹⁹: there are several spurious Sm–Nd isochrons in published literature (summarized in ref. 19). Unpublished Sm–Nd data²⁰ from the Drumbeg layered mafic-ultramafic complex yield an isochron date of $2,910 \pm 55$ Myr, which is similar to the Hamilton *et al.* date¹; the mafic samples with high $^{147}\text{Sm}/^{144}\text{Nd}$ ratios could dominate the $2,920 \pm 50$ Myr isochron, which may be of little, or no, relevance to the grey gneisses.

* Present address: Branch of Isotope Geology, US Geological Survey, MS937, 345 Middlefield Road, Menlo Park, California 94025, USA

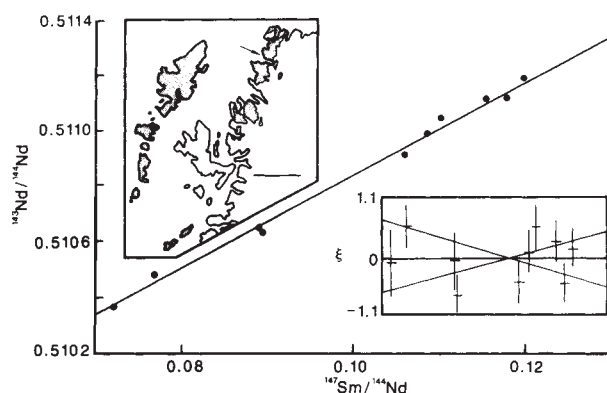


Fig. 1 $^{143}\text{Nd}/^{144}\text{Nd}$ - $^{147}\text{Sm}/^{144}\text{Nd}$ diagram for the central region granulite facies tonalitic gneiss data shown in Table 1. Inset map shows the extent of the foreland Lewisian complex in NW Scotland (scale bar 50 km), and the approximate location of the study area (arrowed). Inset diagram shows deviation of points (ξ) from the best-fit line in parts per 10^4 , and 2σ maximum and minimum slopes ($^{147}\text{Sm}/^{144}\text{Nd}$ range as for main diagram, 2σ error bars on ξ).

The data reported here form part of a regional isotopic study to investigate Lewisian CADS chronology. This study applies available isotopic dating techniques (whole-rock Pb/Pb, Rb-Sr, Sm-Nd) to suites of cogenetic samples, following the rationale outlined by Moorbath and Taylor¹⁹. Ten samples of granulite and retrogressed granulite facies tonalitic gneisses have been collected from the Scourie-Kylesku-Drumbeg area of the central region. Mineralogically, the granulite facies samples comprise the assemblage quartz-plagioclase-clinopyroxene-orthopyroxene, with sieve-textured hornblende-quartz aggregates replacing pyroxene in the retrogressed samples. There is good evidence for the immobility of REE during metasomatic and lower-grade metamorphic events²¹, so that retrogressed samples may be combined with the more pristine granulites. These samples have been analysed for Sm and Nd contents and $^{143}\text{Nd}/^{144}\text{Nd}$ ratios. Table 1 presents sample localities and analytical results.

An errorchron (mean square of weighted deviates, MSWD = 5.7) corresponding to a date of $2,600 \pm 155$ Myr (2σ) with an initial $^{143}\text{Nd}/^{144}\text{Nd}$ ratio of 0.50914 ± 0.00011 is obtained from the data in Table 1 (Fig. 1). $\epsilon_{\text{Nd}}(t)$ for this regression is -2.4 ± 1.9 (calculated at the chondritic uniform reservoir (CHUR) value for $^{147}\text{Sm}/^{144}\text{Nd} = 0.1966$, following the method of ref. 22). Plotted in $\epsilon_{\text{Nd}}(t)$ - t space (Fig. 2), the error polygon for this regression occupies a more light-REE-enriched field than depleted mantle (DM: for details of model illustrated see caption to Table 1) and CHUR reservoirs. Regression data from three

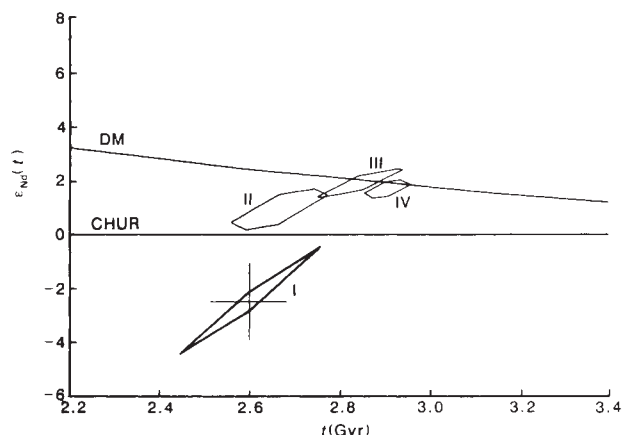


Fig. 2 Nd-isotope evolution diagram ($\epsilon_{\text{Nd}}(t)$ - t) showing the 2σ error polygon for the regression data of Fig. 1 (labelled I) in relation to CHUR and DM model reservoir evolution trends. Also shown are error polygons for Sm-Nd regressions obtained from three mantle-derived layered mafic-ultramafic bodies within the Lewisian (II, Scouriemore: $2,670 \pm 110$ Myr, $\epsilon_{\text{Nd}}(t) = +0.98 \pm 0.73$; III, Achiltibuie: $2,840 \pm 95$ Myr, $\epsilon_{\text{Nd}}(t) = +1.93 \pm 0.52$; IV, Drumbeg: $2,910 \pm 55$ Myr, $\epsilon_{\text{Nd}}(t) = +1.70 \pm 0.32$; ref. 20) illustrating the light-REE-depleted nature of the late-Archaeon sub-Lewisian mantle.

layered mafic-ultramafic complexes in the Lewisian²⁰ indicate that contemporaneous sub-Lewisian mantle closely resembles the DM model; derivation of the gneisses from such a source would require a period of pre-metamorphic crustal residence to develop their light-REE-enriched signature at 2,600 Myr. The observation of negative $\epsilon_{\text{Nd}}(t)$, together with a date that overlaps estimates for the timing of Badcallian granulite facies metamorphism, suggests that the gneisses were subject to regional Nd-isotopic homogenization during this event, a process requiring Nd mobilization.

This inference of Nd-isotopic homogenization and mobilization and, given the geochemical similarity of all the REE, general REE mobility, introduces the possibility of Sm/Nd fractionation. Depleted mantle Sm-Nd model ages (t_{DM}) shown in Table 1, calculated assuming no Sm/Nd fractionation, range from $\sim 2,840$ to $\sim 3,040$ Myr with a terrain average value of $\sim 2,930$ Myr. This large range may be an artefact of Badcallian Nd-isotopic homogenization (Fig. 3). The extent of Sm/Nd fractionation is difficult to assess. Nelson and DePaolo²³ estimate the error in crust formation age t_1 produced by fractionation at t_2 using error $t_1 \approx \Delta f_{\text{Sm}/\text{Nd}}/(t_1 - t_2)$, where $f_{\text{Sm}/\text{Nd}} = ((^{147}\text{Sm}/^{144}\text{Nd})/(^{147}\text{Sm}/^{144}\text{Nd})_{\text{CHUR}} - 1)$. On the basis of partition coefficient data, they consider that the maximum fractionation in Sm/Nd during high-grade metamorphism corresponds

Table 1 Sample localities and Sm-Nd analytical data

Sample	Grid ref.	Sm (p.p.m.)	Nd (p.p.m.)	$^{147}\text{Sm}/^{144}\text{Nd}$	$^{143}\text{Nd}/^{144}\text{Nd} \pm 2\sigma$	$\epsilon_{\text{Nd}}(0)$	$t_{\text{CHUR}} \pm 2\sigma$	t_{DM}
MJW12	NC158435	3.617	18.90	0.1157	0.511147 ± 0.000021	-29.1	$2,790 \pm 40$	2,960
MJW17	NC163430	5.090	28.24	0.1089	0.511020 ± 0.000019	-31.5	$2,800 \pm 30$	2,950
MJW19	NC164430	2.143	16.86	0.0768	0.510490 ± 0.000020	-41.9	$2,720 \pm 30$	2,840
MJW20	NC165417	1.961	13.22	0.0897	0.510646 ± 0.000020	-38.8	$2,820 \pm 30$	2,950
MJW22	NC171407	4.695	23.60	0.1203	0.511218 ± 0.000020	-27.7	$2,820 \pm 40$	2,990
MJW25	NC172397	3.519	20.02	0.1063	0.510942 ± 0.000020	-33.1	$2,840 \pm 30$	2,990
MJW27	NC181386	4.050	20.71	0.1182	0.511148 ± 0.000020	-29.0	$2,880 \pm 40$	3,040
MJW28	NC186377	4.342	23.72	0.1106	0.511072 ± 0.000020	-30.5	$2,760 \pm 40$	2,920
MJW37	NC206356	1.770	11.98	0.0893	0.510671 ± 0.000029	-38.4	$2,780 \pm 40$	2,910
MJW69	NC180327	4.165	34.86	0.0722	0.510376 ± 0.000032	-44.1	$2,750 \pm 40$	2,870

Sm and Nd were separated from a mixture of whole-rock powder and ^{149}Sm - ^{150}Nd spike using cation exchange followed by reversed phase chromatography. Isotopic analyses were performed on a VG Isomass 54E mass spectrometer. Nd-isotopic ratios were normalized to $^{146}\text{Nd}/^{144}\text{Nd} = 0.7219$. $\epsilon_{\text{Nd}}(t) = ((^{143}\text{Nd}/^{144}\text{Nd})_{\text{sample}}/(^{143}\text{Nd}/^{144}\text{Nd})_{\text{CHUR}} - 1) \times 10^4$. t_{DM} model ages were calculated using the algorithm for DM: $\epsilon_{\text{Nd}}(t) = 0.25t^2 - 3t + 8.5$ (ref. 30). Errors on t_{DM} (analytical) are similar to quoted errors on t_{CHUR} .

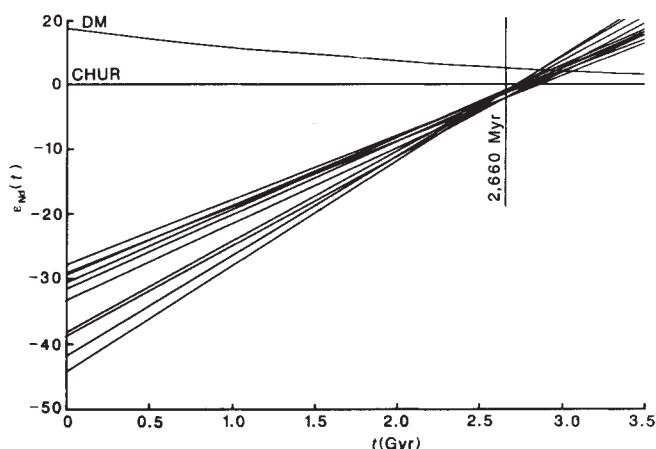


Fig. 3 Nd-isotopic evolution diagram showing individual sample evolution lines converging at ~2,660 Myr. Note the wide range of t_{DM} model ages (intersection with DM model evolution curve) resulting from this convergence.

to $\Delta f_{Sm/Nd} = 0.2$. For the Lewisian, with a probable $t_1 - t_2 < 300$ Myr, this would introduce a maximum uncertainty of ± 60 Myr in model age estimates. Thus the model age data may still be used to suggest ~200–300-Myr crustal residence for the grey gneisses before granulite facies metamorphism at ~2,660 Myr.

The geochronological implication of this study, which indicates a >200-Myr duration for the Lewisian CADS, is similar to that from previously published Sm–Nd data¹. But by considering a more restricted suite of samples, it has been possible to demonstrate Nd-isotopic homogenization in the grey gneisses during Badcallian granulite facies metamorphism. Because Nd-isotopic homogenization implies Nd mobility, this suggests that a mechanism for general REE mobilization is required. A CO₂-rich fluid may be appropriate for effecting REE mobility¹⁶, although there have been objections⁶ to this on the grounds that graphite precipitation should occur at low f_{O_2} . Equally, a melt phase is possible, in accord with latest P – T – a_{H_2O} estimates^{24,25} which indicate that, in some lithologies at least, anatexis may be important. There are other trace-element objections to widespread partial melting, however, such as the similarity of granulite and amphibolite facies Ba/Sr ratios²⁶. The present study cannot constrain this mechanism, although the indication of REE mobility suggests that a CO₂-rich fluid and/or melt phase cannot be discounted. This should be taken into account in future work on the nature of the granulite facies event in the Lewisian complex.

This study provides another example of REE mobility during high-grade metamorphism, previously reported from Antarctica^{27,28} and central Australia²⁹. Given the possibility of Sm/Nd fractionation, such examples cast doubt upon the interpretation of Sm–Nd model ages from granulites as the time of protolith formation, particularly where this precedes metamorphism by a considerable length of time.

I thank Drs S. Moorbath and P. N. Taylor for constructive criticism. Receipt of a NERC research studentship is acknowledged.

Received 10 July; accepted 27 November 1987.

- Hamilton, P. J., Evensen, N. M., O'Nions, R. K. & Tarney, J. *Nature* **277**, 25–28 (1979).
- Pidgeon, R. T. & Bowes, D. R. *Geol. Mag.* **109**, 247–258 (1972).
- Chapman, H. J. & Moorbath, S. *Nature* **268**, 41–42 (1978).
- Moorbath, S. & Taylor, P. N. in *Precambrian Plate Tectonics* (ed. Kroner, A.) 491–525 (Elsevier, Amsterdam, 1981).
- Sheraton, J. W., Skinner, A. C. & Tarney, J. in *The Precambrian of Scotland and Related Rocks of Greenland* (eds Park, R. G. & Tarney, J.) 13–30 (Department of Geology, University of Keele, 1973).
- Fowler, M. B. *J. metamorph. Geol.* **4**, 345–359 (1986).
- Rudnick, R. L., McClelland, S. M. & Taylor, S. R. *Geochim. cosmochim. Acta* **49**, 1645–1655 (1985).

- Tarney, J. & Weaver, B. L. in *Evolution of the Lewisian and Comparable Precambrian High-grade Terrains* (eds Park, R. G. & Tarney, J.) 45–56 (Geol. Soc. spec. Publ. 27, 1987).
- Lyon, T. D. B., Pidgeon, R. T., Bowes, D. R. & Hopgood, A. M. *J. geol. Soc. Lond.* **129**, 389–404 (1973).
- van Breemen, O., Aftalion, M. & Pidgeon, R. T. *Scott. J. Geol.* **7**, 139–152 (1971).
- Moorbath, S., Powell, J. L. & Taylor, P. N. *J. geol. Soc. Lond.* **131**, 213–222 (1975).
- Green, T. H., Brunfelt, A. O. & Heier, K. S. *Earth planet. Sci. Lett.* **7**, 93–98 (1969).
- Hamilton, P. J., O'Nions, R. K. & Evensen, N. M. *Earth planet. Sci. Lett.* **36**, 263–268 (1977).
- McCulloch, M. T. & Wasserburg, G. J. *Science* **200**, 1003–1005 (1978).
- Weaver, B. L. & Tarney, J. *Earth planet. Sci. Lett.* **55**, 171–180 (1981).
- Newton, R. C., Smith, J. V. & Windley, B. F. *Nature* **288**, 45–50 (1980).
- Wendlandt, R. F. & Harrison, W. J. *Contr. Miner. Petrol.* **69**, 409–419 (1979).
- Pride, C. & Muecke, G. K. *Contr. Miner. Petrol.* **73**, 403–412 (1980).
- Moorbath, S. & Taylor, P. N. in *The Chronology of the Geological Record* (ed. Snelling, N. J.) 10–28 (Geol. Soc. mem. 10, 1985).
- Whitehouse, M. J. thesis, Univ. Oxford (1987).
- Humphris, S. E. in *Rare-Earth Element Geochemistry* (ed. Henderson, P.) 317–342 (Elsevier, Amsterdam, 1984).
- Fletcher, I. R. & Rosman, K. J. R. *Geochim. cosmochim. Acta* **46**, 1983–1987 (1982).
- Nelson, B. K. & DePaolo, D. J. *Geol. Soc. Am. Bull.* **96**, 746–754 (1985).
- Barnicoat, A. C. in *Evolution of the Lewisian and comparable Precambrian high-grade terrains* (eds Park, R. G. & Tarney, J.) 73–79 (Geol. Soc. spec. Publ. 27, 1985).
- Cartwright, I. & Barnicoat, A. C. *J. metamorph. Geol.* **3**, 79–99 (1986).
- Tarney, J. & Windley, B. F. *J. geol. Soc. Lond.* **134**, 153–172 (1977).
- DePaolo, D. J., Manton, W. I., Grew, E. S. & Halpern, M. *Nature* **298**, 614–618 (1982).
- Black, L. P. & McCulloch, M. T. *Earth planet. Sci. Lett.* **82**, 15–24 (1987).
- Windrum, D. P., McCulloch, M. T., Chappell, B. W. & Cameron, W. E. *Earth planet. Sci. Lett.* **70**, 27–39 (1984).
- DePaolo, D. J. *Nature* **291**, 193–196 (1981).

Pleistocene dates for the human occupation of New Ireland, northern Melanesia

Jim Allen*, Chris Gosden*, Rhys Jones† & J. Peter White‡

* Department of Archaeology, La Trobe University, Melbourne, Australia

† Department of Prehistory, Australian National University, Canberra, Australia

‡ Department of Anthropology, University of Sydney, Sydney, Australia

Pleistocene dates from three cave sites indicate the human capacity to colonise across two oceanic straits to the east of a former Tasmania–Australia–New Guinea continent by 33 kyr BP. The sites demonstrate exploitation of coastal marine and lowland tropical forest resources. They extend Pleistocene occupation into island Melanesia and demonstrate that the large islands of northern Melanesia have an antiquity of human occupation of the same order as the adjacent Greater Australian continent.

The oldest secure date for human occupation in Greater Australia is now 40 kyr BP¹ for several stone artefacts considered to be in a primary depositional situation between tephras on the uplifted coral terraces of the Huon Peninsula in eastern New Guinea (Fig. 1). Marginally younger dates are claimed from both the southeast and southwest of Australia^{2–3}, while dates in the New Guinea highlands approach 30 kyr BP⁴.

East of New Guinea, the large islands of New Britain and New Ireland were kept separate during the Pleistocene by narrow but deep water barriers. Until now a date of 11 kyr BP from Misisil Cave in southwest New Britain⁵ was the only indication that humans had crossed the first of these barriers by the terminal Pleistocene.

In 1985, as part of the Lapita Homeland Project⁶, excavations were carried out in a series of caves in uplifted coral limestone in northern and central New Ireland. The three sites reported here all have 1–2 m of stratified deposits, containing stone tools, bone and marine shell food debris, hearths and pits. Radiocarbon assays from these sequences (Table 1) indicate human occupation extending back to ~14 kyr BP at Balof 2, ~15 kyr at Panakiwuk and ~33 kyr at Matenkupkum.

Matenkupkum is a cave 50 m from the present shore and 15 m above it, containing 1.4 m of cultural deposits overlying sterile

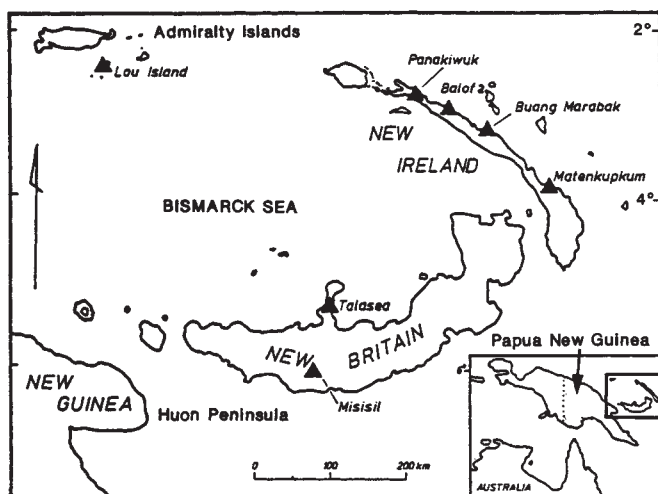


Fig. 1 Bismarck Archipelago showing archaeological sites and other principal locations mentioned in the text.

beach sand and bedrock. Of the seven cultural layers distinguished, the earliest (Layer 7) is dense marine shell midden (Fig. 2). More than 200 flaked stone tools were also recovered from this layer as well as terrestrial animal bones. Its sandy brown matrix was easily distinguished from the sterile beach sand below, leaving no doubt that the earliest deposit above the beach sand is an intact occupation of human derivation. The shells which provided the dating material were part of this human event.

All Matenkupkum shell dates used a single species of gastropod, *Turbo argyrostoma*. From Layer 7, ANU-5070 was divided into degraded (possibly burnt) and undegraded fractions which produced very similar results; two other determinations (ANU-5065 and ANU-5469), 6 m away from the first sample, yielded very similar results. Three of these four determinations overlap with each other at one standard deviation. The exception (ANU-5469) does so at two standard deviations. All fall comfortably within the maximum range of 30–36 kyr BP. Above these, the single Layer 6 date (ANU-5953) confirms the impression of intermittent occupation and gradual accumulation of these lowest deposits gained in the field. The two Layer 5 dates (ANU-5951 and ANU-5952) taken from immediately above ANU-5953 indicate that there was minimal use of the site during the height of the last glacial maximum. Layers 4, 3 and 2 contain the greatest concentrations of bone, stone and shell

discard and are dated by ANU-5467 and ANU-5468 to the terminal part of the Pleistocene. It appears that there was little human use of Matenkupkum during the Holocene and the thin deposits constituting Layer 1 contain glass and metal artefacts relating to Japanese occupation of the cave during the second world war. Two charcoal samples (ANU-5066 and ANU-5067) from Layer 1 yielded concordant modern results. Two further charcoal dates from Layers 3 and 4 (ANU-5068 and ANU-5069) did, however, produce anomalous modern results. These are under further investigation but we currently reject them on other archaeological evidence. Matenkupkum contains no pottery, no *Rattus exulans* remains and no Lou Island obsidian from the Admiralties group. These three data sets occur very commonly throughout the region in sites younger than ~3 kyr BP and all occur, for example, in the upper levels of the two other New Ireland cave sites discussed here.

Continued coastal and marine foraging would have been facilitated by the steeply sloping shoreline in front of the cave. Around 30 kyr BP sea levels were ~–50 m compared with the present, dropping to a minimum –150 m at the last glacial maximum. Rising temperatures after 15 kyr saw sea levels approaching present levels ~10 kyr BP, finally stabilizing at ~6 kyr BP^{7,8}. Although these fluctuations must have affected reef and shell-bed formation, they would not have greatly increased the distance between the cave and the sea. In general, the same is true of the Balof 2 shelter, which although now ~1 km from the coast, also contains marine shells, stone tools and animal bones, scattered through its Pleistocene layers. This shelter contains 1.8 m of cultural deposits overlying sterile clay. The lowest 0.5 m of deposition is contained in a clayey matrix with lenses of bat guano; above this the matrix is greyer and homogenized by human and other taphonomic agents. A hearth dug into the basal sterile clay provided the charcoal for ANU-4848. ANU-4973 and SUA-2502 given an acceptable date for the top of the clayey unit which coincides with the approximate end of the Pleistocene. Although the Pleistocene occupation of Balof 2 may have been episodic, the amounts of human discard are not small. This point is of particular interest because in 1969 Balof 1, a shelter only a few metres from Balof 2 and in the same doline, yielded a basal date of $6,800 \pm 400$ BP (NSW-95)⁹. Thus, although both sites were occupied in the Holocene only Balof 2 has a Pleistocene component.

The third site, Panakiwuk, is situated in a limestone doline ~4 km from the coast. Here, 1.6 m of cultural deposits overlie sterile basal clays. Apparent roof fall has complicated the stratigraphy of the Pleistocene deposits. Human occupation, however, is demonstrated by the presence of stone tools and faunal food debris in the layer preceding that dated by RIDDLE-

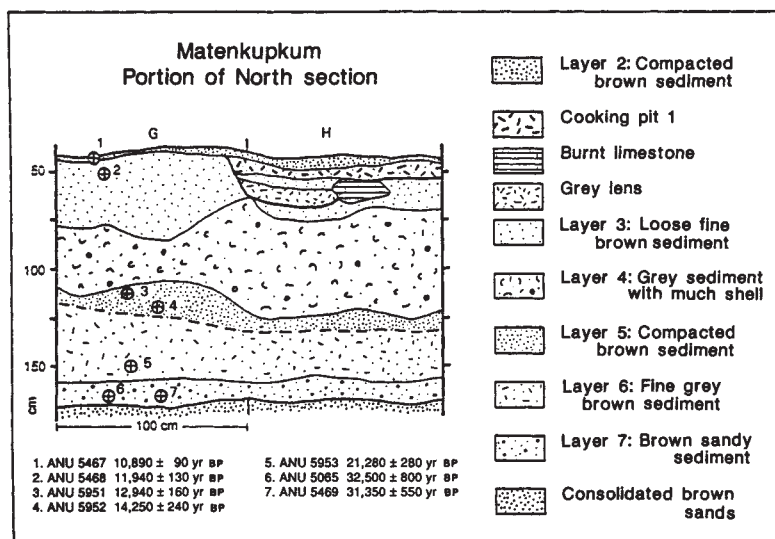


Fig. 2 Stratigraphic section of Matenkupkum, squares G and H, showing the locations of seven of the radiocarbon dates reported in this paper.

Table 1 Radiocarbon dates for Matenkupkum, Balof 2 and Panakiwuk

Matenkupkum Layer	Material	Date BP	Laboratory number
1	Charcoal	Modern	ANU-5066
1	Charcoal	Modern	ANU-5067
2	Shell	10,890 ± 90	ANU-5467
3	Charcoal	Modern	ANU-5068
4	Charcoal	Modern	ANU-5069
4	Shell	11,940 ± 130	ANU-5468
5	Shell	12,940 ± 160	ANU-5951
5	Shell	14,250 ± 240	ANU-5952
6	Shell	21,280 ± 280	ANU-5953
7	Shell	31,350 ± 550	ANU-5469
7	Shell	32,500 ± 800	ANU-5065
7	Shell (degraded)	33,300 ± 950	ANU-5070
7	Shell (undegraded)	32,700 ± 1,550	ANU-5070

Balof 2 Depth below surface (cm)	Material	Date BP	Laboratory number
45-52	Charcoal	3,120 ± 190	ANU-4972
70-80	Charcoal	7,680 ± 510	ANU-4849
122-130	Charcoal	10,560 ± 230	SUA-2502
142-148	Charcoal	9,970 ± 390	ANU-4973
170-174	Charcoal	14,240 ± 400	ANU-4848

Panakiwuk Stratigraphic unit	Material	Date BP	Laboratory number
1	Charcoal	640 ± 130	ANU-5529
2	Charcoal	1,040 ± 110	ANU-5530
2	Charcoal	1,170 ± 70	ANU-5376
4	Charcoal	1,630 ± 130	ANU-5531
6	Bone	8,910 ± 690	ANU-5547
6	Bone	10,160 ± 390	ANU-5546
8	Bone	8,480 ± 500	ANU-5545
8	Bone	10,300 ± 310	ANU-5544
9	Bone	6,780 ± 1,220	ANU-5542
9	Bone	8,000 ± 830	ANU-5541
9	Bone	8,530 ± 520	ANU-5540
11	Charcoal	12,930 ± 210	RIDDL-316 (ANU-4978)
13	Charcoal	15,140 ± 160	RIDDL-531

The dates quoted here are in radiocarbon years using the Libby half life of 5,568 years. The marine shell dates have not been corrected for oceanic reservoir effect; in Australia the accepted correction is -450 ± 35 BP. All bone dates relate to the apatite fraction of the bone. All dates have been corrected for $^{12}\text{C}/^{13}\text{C}$ ratios.

531. Both this determination and RIDDL-316 are accelerator-mass-spectrometry dates on small amounts of charcoal from hearths. Human use of the site during this period, although clearly demonstrated, appears quite episodic. Successive layers indicate a more systematic use of the site into the early Holocene. The suite of dates listed in Table 1 which relate to this period (Stratigraphic Units 6, 8 and 9) contain some anomalies, possibly related to mixing in the only available dating medium in these layers, hundreds of small rat bones. A number of these dates are considered to be minimal because of the possible presence of younger carbonates. Panakiwuk appears to have been abandoned ~ 8 kyr BP and not reused until ~ 2 kyr BP.

These results treble the known time scale for humans in the northern Melanesian islands. They not only relate the initial colonization of these islands to the initial colonization of Greater Australia, but also suggest that the Solomon Islands to the southeast may have been colonized during the Pleistocene as well, as no greater water crossings were involved than those already traversed to reach New Ireland, or indeed to cross Wallacea to Greater Australia.

The second import of the dates presented here is that although marine and coastal subsistence patterns of late Pleistocene

hunter-gatherers are often postulated, higher Holocene sea levels have usually obliterated the evidence. Older marine shell middens are known in both southern and northern Africa¹⁰⁻¹², but those spanning the final 20 kyr of the Pleistocene are infrequent. The suspected rapid uplift of the New Ireland east coast and its plunging profile below present sea levels have preserved in Balof 2 and Matenkupkum important evidence of the human exploitation of these coastal resources during this crucial period of the terminal Pleistocene.

We thank our parent institutions and the National Geographic Foundation for support. Permission to excavate was granted by the New Ireland Provincial Government, the Papua New Guinea Government, Epelis Kasino of Mangai, Sanaile of Medina and Esekia Tomon of Hilalon, Big Men and traditional owners of the sites. Rudy Frank drew the figures.

Received 18 December 1987; accepted 13 January 1988.

- Groube, L., Chappell, J., Muke, J. & Price, D. *Nature* **324**, 453-455 (1986).
- White, J. P. & O'Connell, J. F. *A Prehistory of Australia, New Guinea and Sahul* (Academic, Sydney, 1982).
- Pearce, R. H. & Barbetti, M. *Archaeol. Oceania* **16**, 173-178 (1981).
- Gillieson, D. S. & Mountain, M. J. *Archaeol. Oceania* **18**, 53-62 (1983).
- Specht, J., Lilley, I. & Normu, J. *Austr. Archaeol.* **12**, 13-15 (1981).
- Allen, J. J. *pacif. Hist.* **14**, 186-201 (1984).
- Chappell, J. & Thom, B. G. in *Sunda and Sahul* (eds Allen, J., Golson, J. & Jones, R.) 275-291 (Academic, London, 1977).
- Hope, G. S., Golson, J. & Allen, J. *J. hum. Evol.* **12**, 37-60 (1983).
- Downie, J. E. & White, J. P. *Rec. Austr. Mus.* **31**, 763-802 (1979).
- Klein, R. G. *Science* **197**, 121-126 (1977).
- Singer, R. & Wymer, J. *The Middle Stone Age at Klasies River Mouth in South Africa* (University of Chicago Press, Chicago, 1982).
- McBurney, C. B. M. *The Haua Fteah (Cyrenaica) and the Stone Age of the South-East Mediterranean* (Cambridge University Press, Cambridge, 1967).

Patterns in the distribution and abundance of insect populations

Kevin J. Gaston & John H. Lawton

Department of Biology, University of York, York YO1 5DD, UK

A central problem in ecology is to identify and understand patterns in the distribution and abundance of species^{1,2}. Here, we analyse six patterns for insect populations and explore their inter-relationships. For moths, aphids, carabid beetles, and insects feeding on bracken, we examine links between local population abundance, local population variability, regional distribution, and body size. We show that population characteristics are correlated both theoretically and empirically. Widespread species are generally locally abundant, and have populations that fluctuate more than scarce, geographically restricted species. Predicted effects of body size are less well supported, although common, widespread, widely fluctuating species tend to be small.

We expect the distribution and abundance patterns of species to be influenced by six pairs of relationships (Fig. 1):

- (1) Local population abundance and regional distribution. A positive relationship between population size at a single site and the number of sites a species occupies has been demonstrated for a number of taxonomic groups³⁻⁹, and is expected if widespread species are more flexible in their use of resources^{6,9,10}.
- (2) Local population variability and regional distribution. Species with restricted regional distributions and high local population variability should be rarely encountered, because they have a high probability of extinction¹¹. Thus, we expect a positive correlation between species variability at a site and the number of sites occupied.
- (3) Local population abundance and body size. A negative relationship between these variables is well documented^{12,13}. It could be generated by an increase in *per capita* resource use with increasing body size¹²⁻¹⁴.

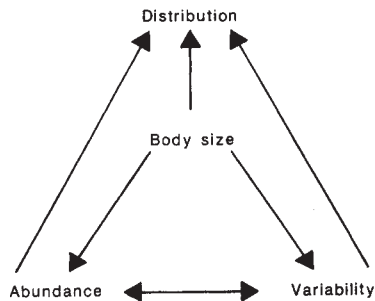


Fig. 1 Theoretical and/or previously reported empirical relationships between population characteristics and body size, with presumed causality indicated by direction of arrows. 'Abundance' and 'variability' refer to local populations; 'distribution' is a measure of the number of local sites occupied by a species throughout a geographical region.

(4) Local population variability and body size. Smaller species have higher intrinsic rates of increase, r , than larger species^{15,16}, which will tend to generate greater deterministic population oscillations¹⁷. Variability may be reinforced if smaller species are more susceptible to environmental disturbances¹⁸. A negative correlation between population variability and body size is predicted, but has not previously been tested, except in comparisons between vertebrates and invertebrates¹⁹.

(5) Regional distribution and body size. Again on the basis that r scales negatively with body size, subsequent to colonization of a new site, smaller species will have faster population growth than larger species. Given that extinction rates are higher for small populations^{20,21}, smaller species should be less vulnerable to extinction upon colonizing a new site than larger species^{22,23}. Thus, we predict a negative relationship between the number of sites occupied and body size. Counter arguments suggest that species of large body size, and hence low population density, will have high probabilities of extinction if they also have small geographic ranges (as total population sizes will be small)^{20,21}. These arguments predict a positive relationship between number of sites occupied and body size⁹.

(6) Local population abundance and local population variability. Means and variances are typically correlated in population data.

We analysed four datasets: 263 species of adult British moths (Lepidoptera, mainly Noctuidae and Geometridae), 97 species

of alate (winged) aphids (Hemiptera, Aphididae), the larval and adult stages of 21 herbivorous insects (Collembola, Diptera, Hemiptera, Hymenoptera and Lepidoptera) feeding upon bracken (*Pteridium aquilinum*), and 63 species of adult Dutch carabids (Coleoptera, Carabidae). Hence, we can test for patterns both within sets of closely related species (moths, aphids or carabids) across habitats and communities, and within a taxonomically disparate assemblage of insects from one community (bracken).

Data for moths and aphids are from light- and suction-trap catches respectively²⁴, with abundance measured as the mean log of the mean number of individuals caught per generation at each of the trapping sites at which individuals were caught (a maximum of 53 sites for moths and 18 for aphids). Local population variability was measured as the mean of the standard deviation of the log of the number of individuals caught at each occupied site^{25,26}. (Alternative measures of variability²⁶ are highly correlated with this measure.) Species distributions were measured very approximately by the number of occupied sites. Body sizes for adult moths were taken from ref. 27; aphid body sizes were very variable within, and covered too narrow a range between, species to justify analysis.

For carabids, population information was available from 1959 to 1966²⁸, analysed as for the moths and aphids, using pitfall catches from four sites with the most complete data. Body sizes were taken from refs 29 and 30; distributions were the number of 10-km squares in which each species occurred³¹.

Population data for bracken insects were obtained for both an open and a woodland site near York between 1980 and 1986^{32,33}. For each species, abundance was measured as the log of the mean of the maximum number of individuals per frond per year in each habitat; variability was the standard deviation of the log of the maximum mean number of individuals per frond across years. Distributions were determined from 216 sites throughout Britain^{32,33}. Body sizes of the feeding-stages were ranked; more quantitative treatment was not justified because of intraspecific variation across life-history stages.

Relationships between local abundance, regional distribution and population variability are particularly clear (Table 1, Fig. 2). With the exception of carabids, widespread species tend to be locally abundant and to have populations that fluctuate more than locally rare, geographically restricted species. It is unclear why carabids appear to be different (Table 1). As the only predominantly carnivorous group they might be expected to show strong correlations between abundance, variability and distribution, if species feeding higher in food chains are more

Fig. 2 Relationships between local population abundance (a-c), local population variability (d-f), and regional distribution for moths, aphids and bracken insects. Data on local abundance for bracken insects are from the woodland site. Statistics are in Table 1. For the moths and aphids, distribution is measured as the number of sites at which each species was caught. For the bracken insects, distribution is the proportion of sites occupied (both open and woodland), this proportion having been corrected for differences in seasonal phenology of the species, such that only those sites sampled in months when species might be present were used. See text for measures of abundance and variability.

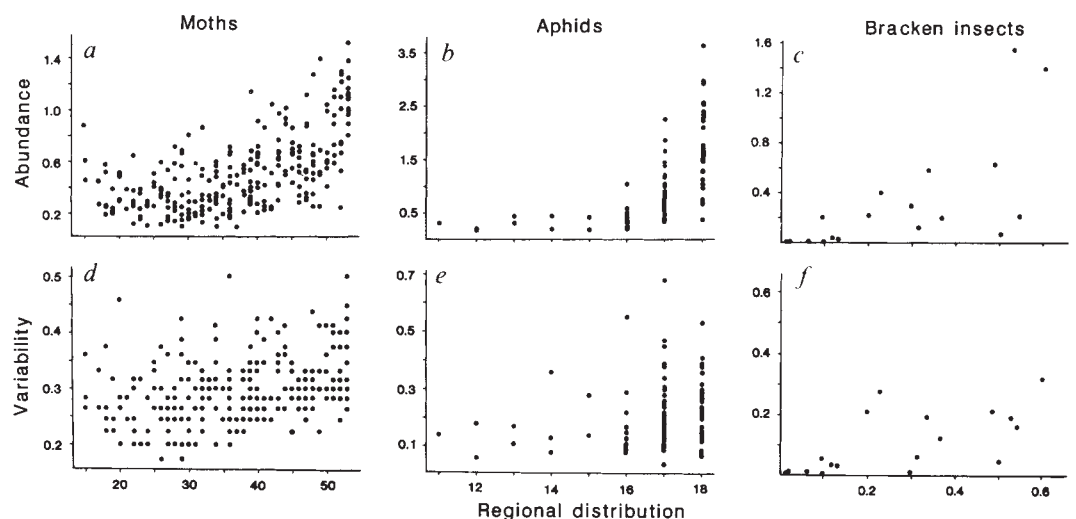


Table 1 Summary of correlation coefficients, and path coefficients⁴³ in parentheses

	N	Abundance distribution	Variability distribution	Variability abundance	Body size abundance	Body size variability	Body size distribution
Moths—all	263	0.593*** (0.616)	0.370*** (-0.034)	0.657***	-0.099 (-0.099)	-0.204*** (-0.204)	0.123* (0.123)
—noctuids	115	0.580***	0.437***	0.738***	-0.079	-0.164	-0.027
—geometrids	109	0.685***	0.370***	0.554***	-0.071	-0.180	0.166
Aphids	97	0.782***	0.244**	0.234*			
Carabids	63	0.096	0.071	0.623***	-0.053	-0.151	0.010
Bracken herbivores							
—open habitat	21	0.501* (0.763)	0.329 (-0.312)	0.840***	-0.616*** (-0.566)	-0.620*** (-0.516)	-0.666*** (-0.671)
—woodland	18	0.463** (-0.016)	0.648** (0.659)	0.726***	-0.605*** (-0.452)	-0.689*** (-0.709)	-0.782*** (-0.596)

* $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$.

vulnerable to stochastic extinctions^{34,35}. It is possible that serious biases in pitfall trap data^{36,37} obscure underlying patterns.

Predicted relationships with body size were more difficult to show statistically (Table 1, Fig. 3). The bracken data are particularly clear (Fig. 3b, d, f), possibly because the insects span a much wider range of body sizes than the other taxa, and because they are the only direct counts of abundance. As measures of population abundance, light, suction and pitfall trap samples are all confounded by other variables³⁶⁻³⁸. Part of the weakness in the body-size relationships may also be real. Not all empirical studies have yielded tight, consistent correlations between population abundance and body size^{9,39-41}, particularly where care has been taken to sample small rare species adequately⁴⁰ that may be missing from literature reviews of population abundance^{12,14,42}. Both predictions about the relationship between size of geographic range and body size are supported by the data. The correlation is positive for moths (but weakly so, and disappears when the single outlying point (Fig. 3e) is removed), and negative (clearly so) for bracken insects. On present evidence we are forced to conclude that body size has an inconsistent effect on population abundance, variation and distribution, and that theoretical expectations are not strongly supported by data from insect populations.

It is difficult to disentangle which variable(s) contribute(s) most to particular population attributes (see Fig. 1). Stepwise multiple-regression analyses do not lead to any simple and consistent conclusions³³. Path analysis⁴³ (Table 1), with regional distribution as the criterion variable and the abundance-variability interaction treated as a simple correlation, suggests that population abundance explains most of the variance for moths and for bracken insects on the open site, whereas variability does so for bracken insects on the woodland site. In all cases, body size explains a large proportion of the variance in distribution, variability and abundance (Table 1).

Our analyses demonstrate some simple empirical and theoretical rules linking population dynamics, distribution and body size. Widespread species are generally locally abundant, and have populations that fluctuate more than scarce, geographically restricted species. For the most reliable data (bracken-feeding insects), within one community, species of small body size were the most widely distributed, and had the most variable and the largest populations; only hints of similar body size relationships exist for other taxa. The generality of these patterns requires testing with data for non-insect taxa, and may lead to a general theory of animal population biology and to an understanding of evolutionary constraints acting on entire species assemblages⁹.

We thank B. Hawkins, S. Juliano, P. Warren, M. Williamson and two referees for their comments. B. McArdle offered statistical advice. K.J.G. is supported by an NERC studentship. Work on bracken was supported by a NERC grant.

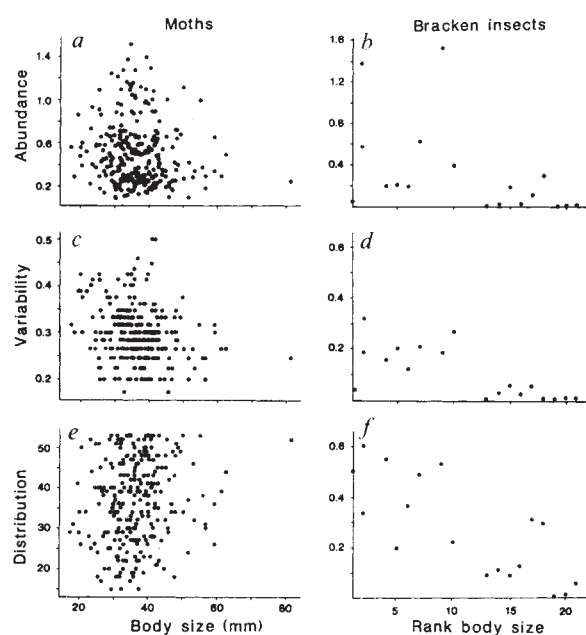


Fig. 3 Relationships of local population abundance (a, b), local population variability (c, d), and regional distribution (e, f), with body size for moths and bracken insects. Local population data for bracken insects is from the woodland site. Statistics are in Table 1. Body size is a continuous variable for the moths (wingspan in mm) and ranked for the bracken insects; see text for details. The positive correlation in e is no longer significant when the single large species is removed from the analysis.

Received 17 September 1987; accepted 13 January 1988.

- McArthur, R. H. *Geographical Ecology: Patterns in the Distribution of Species* (Harper & Row, New York, 1972).
- Brown, J. H. *Am. Zool.* **21**, 877-888 (1981).
- Hanski, I. *Oikos* **38**, 210-221 (1982).
- Fuller, R. J. *Bird Habitats in Britain* (Poyser, Calton, 1982).
- Bock, C. E. & Ricklefs, R. E. *Am. Nat.* **122**, 295-299 (1983).
- Brown, J. H. *Am. Nat.* **124**, 255-279 (1984).
- O'Connor, R. J. & Shrubbs, M. *Farming and Birds* (Cambridge University Press, 1986).
- Bock, C. E. *Ecology* **68**, 124-129 (1987).
- Brown, J. H. & Maurer, B. A. *Am. Nat.* **130**, 1-17 (1987).
- McNaughton, S. J. & Wolf, L. L. *Science* **167**, 131-139 (1970).
- Glazier, D. S. *Oikos* **47**, 309-314 (1986).
- Peters, R. H. *The Ecological Implications of Body Size* (Cambridge University Press, 1983).
- Brown, J. H. & Maurer, B. A. *Nature* **324**, 248-250 (1986).
- Peters, R. H. & Wassenberg, K. *Oecologia* **60**, 89-90 (1983).
- Smith, F. E. in *Dynamics of Growth Processes* (ed. Boell, J.) 274-294 (Princeton University Press, 1954).
- Southwood, T. R. E. in *Theoretical Ecology* (ed. May, R. M.) 30-52 (Blackwell, Oxford, 1981).
- May, R. M. *Theoretical Ecology* (Blackwell, Oxford, 1981).
- Southwood, T. R. E. *J. anim. Ecol.* **46**, 337-365 (1977).

19. Schoener, T. W. in *Community Ecology* (eds Diamond, J. & Case, T. J.) 556–586 (Harper & Row, New York, 1985).
20. MacArthur, R. H. & Wilson, E. O. *The Theory of Island Biogeography* (Princeton University Press, 1967).
21. Leigh, E. G. *J. theor. Biol.* **90**, 213–239 (1981).
22. Lawton, J. H. & Brown, K. C. *Phil. Trans. R. Soc. B* **314**, 607–617 (1986).
23. Crawley, M. J. in *Colonisation, Succession and Stability* (eds Gray, A. J., Crawley, M. J. & Edwards, P. J.) 429–453 (Blackwell, Oxford, 1987).
24. Taylor, L. R. & Woiwod, I. P. *J. anim. Ecol.* **49**, 209–244 (1980).
25. Connell, J. H. & Sousa, W. P. *Am. Nat.* **121**, 789–824 (1983).
26. Williamson, M. *Ecol. Ent.* **9**, 239–241 (1984).
27. Skinner, B. *Colour Identification Guide to Moths of the British Isles* (Penguin, Harmondsworth, 1984).
28. den Boer, P. J. *Misc. Papers* **14** (Landhonwhogeschool, Wageningen, 1977).
29. Joy, N. H. *A Practical Handbook of British Beetles* (Witherby, London, 1932).
30. Lindroth, C. H. *Handbooks for the Identification of British Insects* Vol. IV. Part 2. *Coleoptera: Carabidae* (Royal Entomological Society, London, 1974).
31. Turin, H., Haack, J. & Hengeveld, R. *Atlas of the Carabid Beetles of the Netherlands* (North-Holland, Amsterdam, 1977).
32. MacGarvin, M., Lawton, J. H. & Heads, P. A. *Oikos* **47**, 135–148 (1986).
33. Gaston, K. J. & Lawton, J. H. *Am. Nat.* (in the press).
34. Pimm, S. L. & Lawton, J. H. *Nature* **268**, 329–331 (1977).
35. Fowler, C. W. & MacMahon, J. A. *Am. Nat.* **119**, 480–498 (1982).
36. Southwood, T. R. E. *Ecological Methods: With Particular Reference to the Study of Insect Populations* (Chapman & Hall, London, 1966).
37. Halsall, N. P. & Wratten, S. D. *Ecol. Ent.* **13** (in the press).
38. Wolda, H. *Am. Nat.* **112**, 1017–1045 (1978).
39. Juanes, F. *Am. Nat.* **128**, 921–929 (1986).
40. Morse, D. R., Stork, N. E. & Lawton, J. H. *Ecol. Ent.* **13**, 25–37 (1988).
41. Griffiths, D. *Am. Nat.* **127**, 140–166 (1986).
42. Damuth, J. *Biol. J. Linn. Soc.* **31**, 193–246 (1987).
43. Sokal, R. R. & Rohlf, F. J. *Biometry: The Principles and Practice of Statistics in Biological Research* (Freeman, New York, 1981).

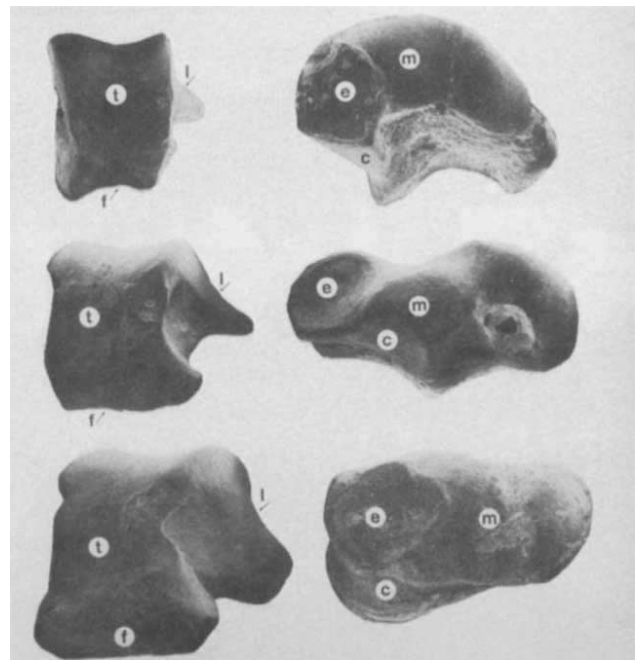


Fig. 1 Posterior view of talus (left), and anterior view of navicular (right). Top row, the anthropoid *Saimiri sciureus*; middle row, the lemuriform *Lemur fulvus*; bottom row, the notharctine adapiforms *Notharctus robustior* (USNM 21968; talus only) and *Cantius ralstoni* (UW unnumbered specimen; navicular only). c, cuboid facet; e, ectocuneiform facet; f, groove for *M. flexor hallucis longus*; l, lateral facet for fibula; m, mesocuneiform facet; t, trochlear articular surface of talus.

Interrelationships among primate higher taxa

K. Christopher Beard*, Marian Dagosto*, Daniel L. Gebo* & Marc Godinot†

* Department of Cell Biology & Anatomy, The Johns Hopkins University, School of Medicine, Baltimore, Maryland 21205, USA
 † Laboratoire de Paléontologie, Université des Sciences et Techniques du Languedoc, CNRS, UA 327, F-34060 Montpellier Cedex, France

One of the most controversial issues in primate palaeontology concerns the phylogenetic position of the extinct primate infraorder Adapiformes^{1,2}. During the Eocene, this group of primates of modern aspect possessed a holarctic distribution^{1–12}, and may have been present in the poorly known Palaeogene of Africa¹⁰. Mainly on the basis of craniodental morphology, at least four hypotheses have been proposed concerning the phylogenetic interrelationships among adapiforms and other primate higher taxa: (1) that adapiforms are ancestral to both lemuriforms (including Lorisoidea) and anthropoids^{8–10,13}; (2) that adapiforms cannot be shown to possess a special phylogenetic relationship with either lemuriforms^{14,15} or anthropoids^{15–18}; (3) that adapiforms are the sister taxon of lemuriforms^{16–18}; and (4) that Adapiformes is not a natural, monophyletic group, but rather consists of nested clades within the radiation of lemuriforms^{19–25}. Here, we describe features of the ankle and wrist joints of several adapiform taxa that provide an independent test of the preceding hypotheses. These traits suggest that lemuriforms are monophyletic with respect to known adapiforms, but that adapiforms nevertheless are their stem lineage (*sensu* Ax²⁶).

Fossil ankle and wrist bones of several adapiform genera representing both of the major subfamilies, Notharctinae and Adapinae, have been reported previously^{27–30}. In addition to these, we have examined numerous previously undescribed fossils, including well-preserved carpals from partial skeletons of the middle Eocene notharctine *Smilodectes*³¹ and newly collected tarsals of the early Eocene notharctine *Cantius*.

In lemuriforms the talofibular joint facet on the talus slopes gently laterally for its entire extent (Fig. 1). With the exception of adapiforms, in other primates, including the archaic

Plesiadapis, this facet is vertical except at its plantar surface where it flares abruptly laterally, resulting in a small, pointed plantar process. All adapiforms for which the talus is known (*Cantius*, *Notharctus*, *Smilodectes*, *Adapis*, *Leptadapis*, *Caenopithecus*, the Messel adapid) exhibit the condition otherwise found only in lemuriforms. The talofibular joint morphology of adapiforms and lemuriforms is unique not only among primates, but among all other therians as well, and probably represents a derived condition. A second important aspect of talar morphology is the position of the groove for the tendon of *M. flexor hallucis longus* on the posterior surface of the bone. This groove lies lateral to the posterior part of the tibiotalar joint in lemuriforms, whereas in other primates and therians it is plantar and central to the facet (Fig. 1). Again, all known adapiforms have the derived morphology otherwise found only in lemuriforms.

Lemuriforms and notharctines are also distinguished from other primates and therians by details of the naviculocuboid articulation (the navicular is unknown in adapines). In the former taxa the naviculocuboid articulation is broad, and as a result the facet for the cuboid on the navicular is contiguous with the facets for both the ectocuneiform and the mesocuneiform (Fig. 1). In other primates and therians the facet for the cuboid on the navicular abuts only against the facet for the ectocuneiform on this bone.

The centrale of the lemuriform midcarpal joint is relatively wide mediolaterally, so that its medial end articulates with the hamate; this never occurs in other living primates³². Also, the embrasure between the capitate and trapezoid for articulation with the centrale is very narrow in lemuriforms, so that the articulation between the centrale and the trapezoid occurs in an oblique plane through the wrist. In other living primates and therians this embrasure is much wider and the articulation between the centrale and the trapezoid lies in a transverse plane (Fig. 2). In both *Adapis* and *Smilodectes*, the only two

Fig. 2 Dorsal view of carpus (top), and proximal view of pisiform (bottom). From left to right, the lemuriforms *Lemur macaco* (dorsal view of wrist only) and *Varecia variegata* (pisiform only); the anthropoid *Saimiri sciureus*; the adapine adapiform *Adapis parisiensis* (RD 311); and the notharctine adapiform *Smilodectes gracilis* (USNM 13251A, right articulated capitale and centrale; USNM 256745, left pisiform). Partial wrist of the *Adapis* specimen is depicted in an oblique dorsal view, as it was originally discovered. c, centrale; t, facet for triquetrum; u, facet for ulnar styloid process.

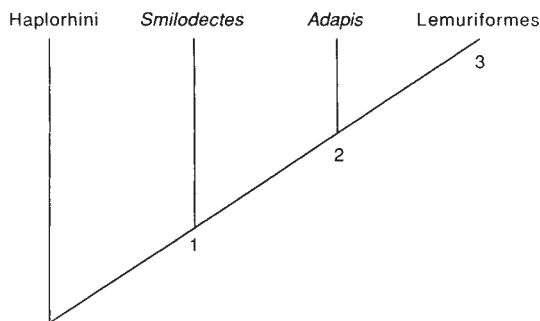
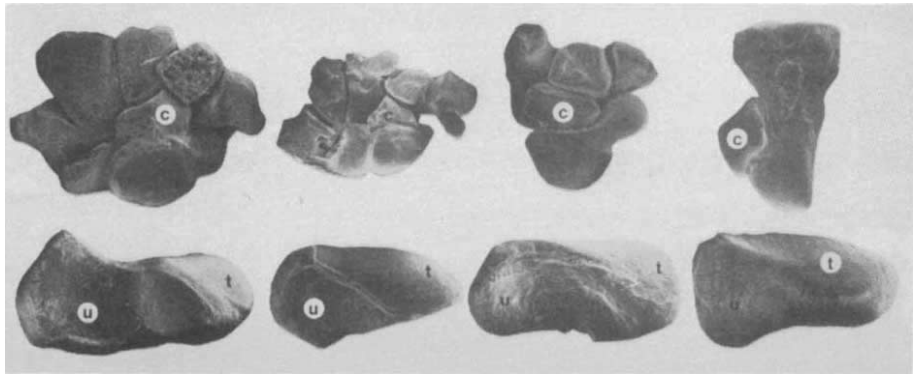


Fig. 3 Cladogram illustrating interrelationships among adapiforms and other primates based on shared derived features of ankle and wrist anatomy. Synapomorphies defining node 1 (Strepsirhini) include: laterally sloped talofibular facet, lateral position of groove for *M. flexor hallucis longus*, and confluence of naviculocuboid and mesocuneiform articular facets on navicular. Synapomorphy defining node 2 (unnamed taxon): facet for ulnar styloid process on pisiform mediolaterally extensive and deeply excavated. Synapomorphy defining node 3 (Lemuriformes): centrale mediolaterally wide, articulating with hamate, centrale-trapezoid articulation in oblique plane.

adapiforms for which the relevant anatomy is known, the structure of the central midcarpal joint is the same as that found in living *Tarsius* and anthropoids, and differs markedly from that found in lemuriforms (Fig. 2).

In lemuriforms the facet for the ulnar styloid process on the pisiform is much more extensive mediolaterally and is much more deeply excavated than is the case in other living primates that have full ulnocarpal contact. The first of these differences is best visualized by comparing the mediolateral extent of the facet for the ulnar styloid process on the pisiform with that of the confluent facet for the triquetrum on this bone (Fig. 2). In lemuriforms the facet for the ulnar styloid process extends more medially than does the facet for the triquetrum, whereas these two facets are equally extensive mediolaterally in other extant primates. In the middle Eocene notharctine *Smilodectes*, the structure of the ulnocarpal articulation is the same as that of living *Tarsius*, platyrrhines and cercopithecoids, and the pisiform lacks the traits described above as characteristic of lemuriforms. In late Eocene *Adapis*, however, the ulnocarpal articulation is the same as that of lemuriforms (Fig. 2). We interpret the ulnocarpal articulation of lemuriforms and *Adapis* as the derived pattern for primates, because of the otherwise ubiquitous occurrence of the alternative character state throughout the order and in other therian mammals.

In summary, the anatomy of the talofibular joint, the naviculocuboid joint, and the position of the groove for the tendon of *M. flexor hallucis longus* all appear to be synapomor-

phies linking Eocene adapiforms with lemuriforms (Fig. 3, node 1). These traits therefore support the monophyly of the primate suborder Strepsirhini, which has long been assumed but never clearly demonstrated^{14,15}. The possible functional implications of these features are discussed elsewhere³³⁻³⁵. The anatomy of the ulnocarpal joint appears to be a synapomorphy linking *Adapis* with lemuriforms to the exclusion of all other living and fossil primates for which the relevant anatomy is known, including the notharctine *Smilodectes*. *Adapis* itself possesses uniquely derived features of incisal and postcranial anatomy which preclude it from being directly ancestral to lemuriforms^{17,18,30}. Nonetheless, we interpret the shared derived occurrence of a similar ulnocarpal articulation in both taxa as indicative of more recent common ancestry between *Adapis* and lemuriforms than between either of these taxa and the notharctine *Smilodectes* (Fig. 3, node 2). Finally, the anatomy of the midcarpal joint appears to be a synapomorphy linking lemuriforms to the exclusion of both *Adapis* and *Smilodectes* (Fig. 3, node 3). This character therefore corroborates evidence from other anatomical regions, most notably the lemuriform toothcomb, which suggests lemuriform monophyly with respect to known adapiforms, contrary to recent proposals based on molar morphology¹⁹⁻²⁵. We conclude that Adapiformes is paraphyletic, and that lemuriform primates probably arose from an as yet unknown Old World adapine relatively late in Palaeogene time, perhaps as recently as the late Eocene.

Received 25 September; accepted 30 December 1987.

1. Szalay, F. S. & Delson, E. *Evolutionary History of the Primates* (Academic, New York, 1979).
2. Hoffstetter, R. *Bull. Mem. Soc. Anthropol. Paris* **4**, 327-346 (1977).
3. Simons, E. L. *Bull. Br. Mus. Nat. Hist., Geol.* **7**, 1-36 (1962).
4. Godinot, M. *C. r. heb. Seanc. Acad. Sci., Paris*, **D286**, 1869-1872 (1978).
5. Gingerich, P. D. & Simons, E. L. *Contr. Mus. Paleont. Univ. Mich.* **24**, 245-279 (1977).
6. Gingerich, P. D. *J. Paleont.* **53**, 153-163 (1979).
7. Gingerich, P. D. *Nature* **319**, 319-321 (1986).
8. Gingerich, P. D. in *Evolutionary Biology of the New World Monkeys and Continental Drift* (eds Ciochon, R. L. & Chiarelli, A. B.) 123-138 (Plenum, New York, 1980).
9. Gingerich, P. D. *Contr. Mus. Paleont. Univ. Mich.* **24**, 163-170 (1975).
10. Gingerich, P. D. *Geobios, Mem. Spec.* **1**, 165-182 (1977).
11. Russell, D. E. & Gingerich, P. D. *C. r. heb. Seanc. Acad. Sci., Paris (Ser. 2)* **304**, 209-214 (1987).
12. Beard, K. C. *Am. J. phys. Anthropol.* (in the press).
13. Gingerich, P. D. & Schoeninger, M. *J. hum. Evol.* **6**, 483-505 (1977).
14. Rasmussen, D. T. *J. hum. Evol.* **15**, 1-12 (1986).
15. Cartmill, M. & Joysey, K. A. in *Recent Advances in Primatology* Vol. 3, *Evolution* (eds Chivers, D. J. & Joysey, K. A.) 205-214 (Academic, London 1978).
16. Rosenberger, A. L. & Szalay, F. S. in *Evolutionary Biology of the New World Monkeys and Continental Drift* (eds Ciochon, R. L. & Chiarelli, A. B.) 139-157 (Plenum, New York, 1980).
17. Rosenberger, A. L., Strasser, E. & Delson, E. *Folia Primatol.* **44**, 15-39 (1985).
18. Rosenberger, A. L. & Strasser, E. *Primates* **26**, 73-84 (1985).
19. Schwartz, J. H., Tattersall, I. & Eldredge, N. *Yb. phys. Anthropol.* **21**, 95-133 (1978).
20. Schwartz, J. H. & Tattersall, I. *Anthrop. Pap. Am. Mus. nat. Hist.* **55**, 271-283 (1979).
21. Schwartz, J. H. & Tattersall, I. *Folia Primatol.* **39**, 178-186 (1982).
22. Schwartz, J. H. & Tattersall, I. *Anthrop. Pap. Am. Mus. nat. Hist.* **57**, 344-352 (1983).
23. Schwartz, J. H. & Tattersall, I. *Anthrop. Pap. Am. Mus. nat. Hist.* **60**, 1-100 (1985).
24. Schwartz, J. H. & Tattersall, I. *J. hum. Evol.* **16**, 23-40 (1987).
25. Schwartz, J. H. in *Comparative Primate Biology. Vol. 1, Systematics, Evolution, and Anatomy* (eds Swindler, D. R. & Erwin, J.) 1-41 (Liss, New York, 1986).
26. Ax, P. *Cladistics* **1**, 279-287 (1985).
27. Godinot, M. & Joffroy, F. K. in *Actes du Symposium Paleontologique G. Cuvier* (eds Buffetaut, E., Mazin, J. M. & Salmon, E.) 221-243 (Montbéliard, 1984).

28. Gregory, W. K. *Mem. Am. Mus. nat. Hist.* 3, 49-243 (1920).
29. Decker, R. L. & Szalay, F. S. in *Primate Locomotion* (ed. Jenkins, F. A., Jr) 261-291 (Academic, New York, 1974).
30. Dagosto, M. *Folia Primatol.* 41, 49-101 (1983).
31. Beard, K. C. & Godinot, M. *J. hum. Evol.* (in the press).
32. Jouffroy, F. K. in *Lemur Biology* (eds Tattersall, I. & Sussman, R. W.) 149-192 (Plenum, New York, 1975).
33. Gebo, D. L. *J. hum. Evol.* 15, 421-430 (1986).
34. Gebo, D. L. *Folia Primatol.* (in the press).
35. Dagosto, M. *J. hum. Evol.* (in the press).

A test of the short-term advantage of sexual reproduction

Steven E. Kelley*§, Janis Antonovics†
& Johanna Schmitt‡

* Department of Botany, Duke University, Durham, North Carolina 27706, USA

† Department of Botany, Duke University, Durham, North Carolina 27706, USA

‡ Graduate Program in Ecology and Evolutionary Biology, Division of Biology and Medicine, Brown University, Providence, Rhode Island 02912, USA

The ubiquity of sex in the face of what is called the 'cost of meiosis' represents one of the major paradoxes of evolutionary biology¹⁻³. Theoretical work has shown that sexual females can only be maintained in mixed populations of sexual and asexual females if their offspring are of the order of twice as fit as those of asexual females¹⁻⁸. Previous experiments using the grass *Anthoxanthum odoratum* have shown substantial fitness advantages for genetically variable arrays of individuals relative to genetically uniform arrays planted in particular experimental designs⁹⁻¹². Yet, in natural conditions progeny do not become arranged at set spacings in arbitrary places within a population, but are dispersed around parents to form a 'seed shadow' with decreasing density away from source¹³. We report here the results of an experiment designed to simulate the natural dispersal pattern of progeny around parents when these progeny are either genetically uniform and identical to parent, or sexually-produced by those parents and genetically variable. Our data show that sexually generated progeny of *Anthoxanthum odoratum* have reproductive rates, summed over two years, that are 1.43 times those of their asexually generated siblings.

Anthoxanthum odoratum, sweet vernal grass, was sampled from a mown field on the Duke University campus, in Durham, North Carolina⁹⁻¹². In May 1982, both single tillers and several inflorescences of *A. odoratum* were taken from plants located at 4-metre intervals in a 6 × 5 sample grid in the central part of the field. At alternating points, 'small adults' (fewer than five inflorescences) and 'large adults' (greater than five inflorescences) were sampled. The median number of inflorescences in the population was 5.1 in 1981. Tillers, sampled from parents, were grown in the glasshouse, and divided to generate 'asexual progeny'. Seeds were germinated in flats in August 1982, grown in the glasshouse, and plants divided to generate 'sexual progeny'. Hence, in this experiment all comparisons were between ramets planted as single tillers. Because *A. odoratum* is self-incompatible and wind-pollinated, individuals deriving from seeds were likely to be highly genetically variable. At five sites, (two large and three small adults), an insufficient number of either sexual or asexual offspring were obtained through cloning. Therefore, 25 parents were included in the experiment.

Progeny were planted with minimal disturbance of the vegetation in arrays which simulated the natural pattern of dispersal about their parents (Fig. 1). Sexual and asexual progeny were identically marked with plastic-coated wire and a plastic tag.

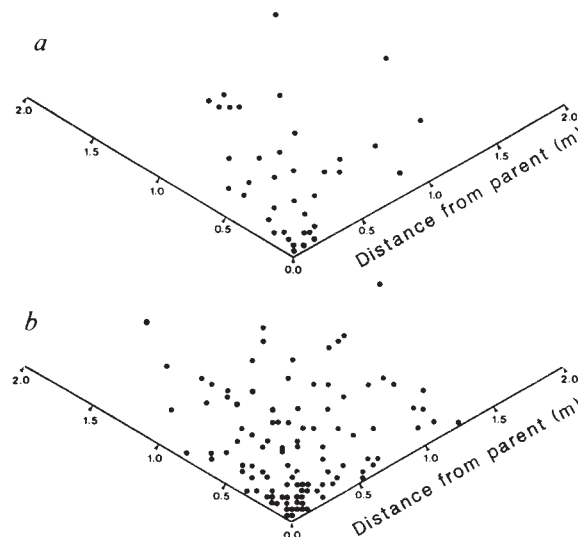


Figure 1 Sample patterns of progeny dispersal for small (a) and large (b) parents used in our experiment. The identical patterns were used as a template to plant an arc of sexual progeny on one side of the parent, and a separate identical arc of asexual progeny on the other side of the parent. Separate simulated patterns of progeny dispersal in an arc of 120° were generated for each parent using a computer, by approximating the measured distance of seed dispersal with a randomly-generated normal distribution of mean 0 m and standard deviation 0.71 m, and generating for each seed a random arc position, repeating this process until the appropriate number of seed positions was obtained. The measured distance of seed dispersal was obtained by surrounding each of eight adults with three plastic sheets (two measuring 1.0 × 1.9 m, one measuring 1.0 × 4.0 m) covered with Tanglefoot, a sticky substance (Tanglefoot Company, Grand Rapids, Michigan) so that dispersed seeds could be collected and their location noted. The appropriate number of progeny per parent was calculated as: number of seeds per parent (large or small) × frequency of germination × survivorship from germination to the end of first year. The number of seeds per parent was estimated by counting the number of inflorescences of every individual within a 1 × 5 m area in the center of the field. Mean number of seeds per inflorescence was determined by randomly subsampling inflorescences and counting the number of spikelets (there is a single seed per spikelet). Germination was determined by planting seeds directly in the field. Survivorship of seedlings for the first year subsequent to germination was measured by monitoring all naturally occurring seedlings found within five randomly located 0.25 × 1 m quadrants. These methods and results are described in greater detail elsewhere²⁴. The calculations showed 360 one-year old progeny around each large adult and 120 one-year old progeny around each small adult.

The orientation of the arcs of sexual and asexual progeny was random about the parent. During data collection, progeny type was not known so there could be no experimenter bias influencing the results of the experiment. A total of 3,920 ramets were planted in November 1983. Individuals were relocated in May 1984 and 1985, and the number of vegetative and reproductive tillers counted. Fitness was estimated as the number of inflorescences summed over two years. The number of inflorescences is strongly correlated with the total number of spikelets ($R^2 = 0.91$, $P < 0.001$, $N = 46$)⁹ and, therefore, with male and female contribution to fitness (each spikelet has one seed and three anthers).

In this experiment, the relative fitness of sexual reproduction, estimated as the ratio of summed reproductive output of sexual to asexual progeny, calculated for each of the 25 parents, log transformed, and then averaged was 1.43 ± 0.29 . The magnitude of the observed advantage of sexual relative to asexual progeny was significant in an experiment which only measured relative differences in the performance of adult tillers planted in the

§ Present address: NERC Institute of Virology, Mansfield Road, Oxford, OX1 3SR, UK.

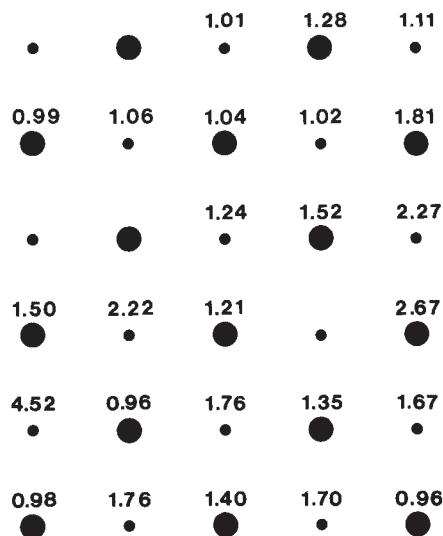


Fig. 2 A plot of the location of parents shown along with the mean relative fitness of sexual versus asexual progeny for each of the parents calculated separately. Large circles represent large parents; small circles represent small parents.

field (Table 1). Evidence from previous experiments indicates that there is also a substantial genetic component to seedling establishment in this population of *A. odoratum*¹¹ and that processes favouring more genetically variable sibships occur during the seedling establishment phase¹². Thus, the advantages observed here for the adult stage of the life cycle may be magnified if earlier seedling stages could be included. The absence of apomictic seed from *A. odoratum* prevents a precise estimate of this relative advantage. Although the measured advantage reported here was less than twofold, 35.6% of the progeny remained alive after two years, and evidence from previous experiments^{9,10} suggests that the difference between sexual and asexual progeny often continues to increase when data from later years are included.

There was considerable variation among parents in the relative fitness (RF) of sexual progeny (Fig. 2; Table 1). RFs varied between 0.96 and 4.52 and were bimodally distributed, with one peak at 1.08 (13 parents), and a second peak at 2.06 (12 parents). The RF of sex was higher for small parents (1.60 ± 0.45) than for large parents (1.30 ± 0.37), but this difference was not statistically significant (test statistic¹⁴ $t' = 0.52$, 22 d.f.). Interactions among siblings were more likely to occur among the progeny of large parents. For small parents 37.4% and for large parents, 53.6% of progeny had at least one sibling within 5 cm. (Single tillers begin to negatively interfere with each others performance within a distance of 5 cm¹⁰.) These results appear to suggest that sib interactions do not enhance the advantage for sexual progeny.

Table 1 Analysis of variance of fitness transformed with a square-root function to improve normality

Source	d.f.	SS	F
Parent	25	209.07	22.04***
Distance	1	1.84	4.85*
Sex	1	34.88	91.95***
Distance*Parent	25	16.69	1.76*
Parent*Sex	25	49.26	5.19***
Distance*Sex	1	0.00	0.00
Distance*Parent*Sex	24	16.20	1.78**
Error	3,817	1,448.05	
Total	3,919	1,775.99	

* $P < 0.05$; ** $P < 0.001$; *** $P < 0.0001$.

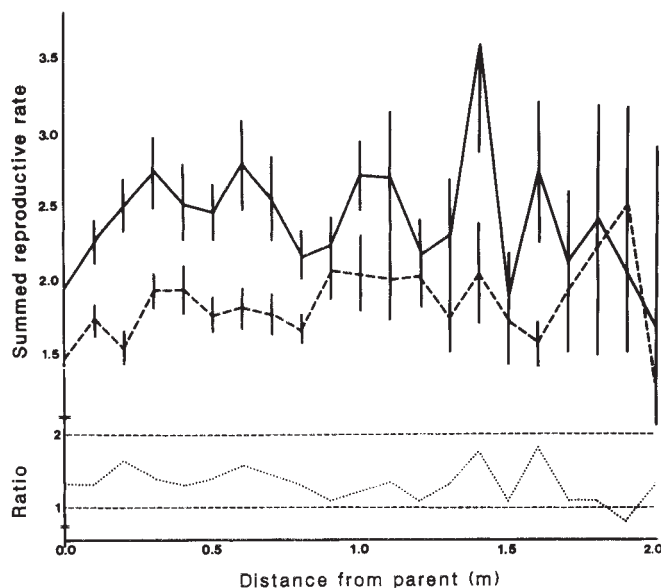


Fig. 3 Mean reproductive rates for sexually (solid line) and asexually (broken line) produced progeny at different distances from the parent measured in terms of inflorescence number. Values shown are ± 1 s.e.m. The ratio of the performance of sexually to asexually produced progeny (a measure of relative fitness) is shown below (dotted line).

The mean performance of sexually and asexually generated progeny in relation to distance from parent provides evidence as to the mechanism(s) maintaining sexual reproduction in the population. The benefit of producing genetically variable progeny have been suggested to derive from: (1) a fluctuating environment¹⁻³, (2) sib competition¹⁵⁻¹⁷, and (3) host-pathogen (predator-prey) interactions¹⁸⁻²³. If a spatially varying environment was the primary selective force favouring sex, then in our experiment the advantage for sexual progeny should have increased with increasing distance from the parent. If sib competition was the principal selective force, then the advantage for sexual progeny should have been greater close to the parent, where siblings were densest (Fig. 1). But our data show no clear pattern of relative advantage (Fig. 3; Table 1). Highly significant differences between the summed reproductive rates of sexually and clonally generated progeny were evident both close to and distant from the parent.

Other experiments in this *Anthoxanthum* population have revealed fitness advantages for minority genotypes not attributable to competition⁹⁻¹²; in one case this minority advantage was clearly linked to infestation by aphids¹². These earlier results suggest that the advantages observed for sexual progeny may result from frequency dependent attack by predators or pathogens. Clearly, more experiments are needed to pinpoint the causes of the effects observed here. But the present study demonstrates that advantages of sexual reproduction, whatever their cause, may be substantial and immediate in natural populations. The essential simplicity of the approach we have used should encourage an examination of these advantages (if they exist) in plants with diverse breeding systems or natural levels of asexual reproduction by vegetative means.

We thank S. Paulsen for field assistance, T. P. Sun for proof-reading the data, and B. C. Clarke for helpful comments on the manuscript.

Received 15 November 1987; accepted 18 January 1988.

1. Williams, G. C. *Sex and Evolution* (Princeton University Press, New Jersey, 1975).
2. Maynard Smith, J. *The Evolution of Sex* (Cambridge University Press, Cambridge, 1978).
3. Bell, G. *The Masterpiece of Nature: The Evolution and Genetics of Sexuality* (University of California Press, Berkeley, 1982).

4. Charlesworth, B. J. *theor. Biol.* **84**, 655-671 (1980).
5. Lloyd, D. G. *Evol. Biol.* **13**, 69-111 (1980).
6. Harper, A. B. *Heredity* **48**, 107-116 (1982).
7. Uyenoyama, M. K. *Evolution* **38**, 87-102 (1985).
8. Uyenoyama, M. K. *Evolution* **39**, 1194-1206 (1985).
9. Antonovics, J. & Ellstrand, N. C. *Evolution* **38**, 103-115 (1984).
10. Ellstrand, N. C. & Antonovics, J. *Evolution* **39**, 657-666 (1985).
11. Schmitt, J. & Antonovics, J. *Evolution* **40**, 830-836 (1986).
12. Schmitt, J. & Antonovics, J. *Evolution* **40**, 837-842 (1986).
13. Levin, D. A. & Kerster, H. W. *Evol. Biol.* **6**, 139-220 (1974).
14. Brownlee, K. A. *Statistical Theory and Methodology* (Krieger, Florida, 1984).
15. Maynard Smith, J. J. *theor. Biol.* **63**, 245-258 (1976).
16. Taylor, P. D. *J. theor. Biol.* **81**, 407-421 (1979).
17. Bulmer, M. K. *J. theor. Biol.* **82**, 335-345 (1980).
18. Jaenike, J. *Evol. Theory* **3**, 191-194 (1978).
19. Bremermann, H. J. *J. theor. Biol.* **87**, 671-702 (1980).
20. Hamilton, W. D. *Oikos* **35**, 282-290 (1980).
21. Tooby, J. *J. theor. Biol.* **97**, 557-576 (1982).
22. Law, R. & Lewis, D. H. *Biol. J. Linn. Soc.* **20**, 249-276 (1983).
23. May, R. M. & Anderson, R. M. *Proc. R. Soc.* **219**, 281-313 (1983).
24. Kelley, S. E. thesis, Duke Univ. (1985).

Primary genetic control of somatic sexual differentiation in a mammal

Wai-Sum O*, R. V. Short, Marilyn B. Renfree & G. Shaw

Departments of Anatomy & Physiology, Monash University, Melbourne, Victoria 3168, Australia

The classical view of mammalian sexual differentiation is that a gene on the Y chromosome transforms the indifferent gonad into a testis¹. The Leydig cells then secrete androgen which stimulates the development of the male reproductive tract, and the Sertoli cells secrete Mullerian inhibitory substance which inhibits the development of the female reproductive tract^{2,3}. In the absence of a testis, the Mullerian duct develops into the Fallopian tubes, uterus and vagina. Thus the whole of sexual differentiation is thought to be hormonally mediated as a consequence of this initial genetic determination of gonadal sex. We have found evidence in a marsupial mammal for extensive sexual dimorphisms which precede any morphological differentiation of the gonads. Thus the classical view of mammalian sexual differentiation may have over-emphasized the role of testicular hormones, and overlooked earlier genetic effects.

Tammar wallabies, *Macropus eugenii*, were used to provide newborn young. Embryonic material was obtained from females at known stages of pregnancy⁴. Fifteen pouch-young were killed by decapitation on the day of birth, sexed by karyotype⁵, and sectioned histologically. The cross-sectional areas of the gonads, Wolffian ducts and Mullerian ducts, including lumina, were measured using an ASM digitizing tablet on every fifth section at $\times 10$ magnification and on every tenth section at $\times 25$ magnification. The volumes of the gonads and the genital ducts were then calculated using Simpson's formula⁶.

There was no significant difference ($P > 0.05$) between the head lengths of males ($\bar{x} = 7.2 \pm \text{s.e.m. } 0.2 \text{ mm}$, $n = 7$) and females ($7.6 \pm 0.2 \text{ mm}$, $n = 8$) on the day of birth. The gonads were still at the undifferentiated stage in all the animals. The gonadal ridges lay medial to the mesonephros and extended from the level of the second lumbar vertebra to the fourth lumbar vertebra; there was no difference in position of the gonads between males and females. Similarly, there was no significant difference ($P > 0.05$) in gonadal volume between males ($0.071 \pm 0.007 \text{ mm}^3$) and females ($0.068 \pm 0.006 \text{ mm}^3$), nor were there any histological differences (Fig. 1a, b). The germ cells were clearly recognizable as large cells with clear cytoplasm, large nuclei

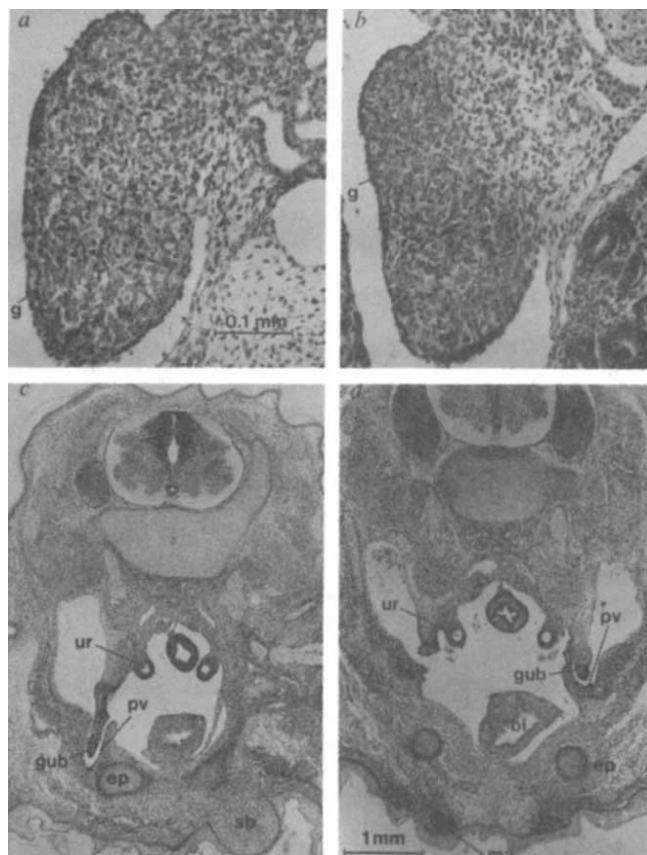


Fig. 1 Histological sections of tammar wallabies on the day of birth. The body was immersion-fixed in neutral buffered formalin for 7 days, embedded in paraffin wax, and transverse serial sections (5 μm) were stained with haematoxylin and eosin. The gonads (g) of males (a) and females (b) are similar. However, males (c) have scrotal bulges (sb) and a long gubernaculum (gub) and processus vaginalis (pv), whereas females (d) have mammary anlagen (ma) and a short gubernaculum and processus vaginalis. Abbreviations: ur, ureters; bl, bladder; ep, epipubic cartilage. Scale bars: a, b, 0.1 mm; c, d, 1.0 mm.

and prominent nucleoli; their distribution between cortex and medulla was identical in males and females.

There was no evidence of any hormone secretion by the male gonads at this stage; the Wolffian duct of males was not significantly larger than that of females ($0.0101 \pm 0.0023 \text{ mm}^3$ versus $0.0082 \pm 0.0008 \text{ mm}^3$, $P > 0.05$), and there was no evidence of reduced Mullerian duct volume in males as compared to females ($0.0012 \pm 0.0001 \text{ mm}^3$ versus $0.0012 \pm 0.0002 \text{ mm}^3$, $P > 0.05$). This is consistent with reports in the rat, rabbit and pig that testicular hormone production only commences after the gonads have differentiated morphologically⁷⁻⁹.

Despite the lack of gonadal differentiation at birth there was already clear evidence of some sexual dimorphism. In all males, scrotal bulges were visible anterior to the phallus (Fig. 1c). These appeared as loose connective tissue swellings, covered by a squamous epithelium slightly thicker than other parts of the body. Scrotal bulges were never seen in females, but mammary anlagen were visible histologically as semi-circular invaginations of the basal layers of the epidermis surrounded by a dense mesenchymal region in the dermis (Fig. 1d). Lateral to the mammary anlagen were two mesodermal ridges, which were presumed to be pouch anlagen. Neither mammary nor pouch anlagen were ever seen in males. The gubernaculum and processus vaginalis were also sexually dimorphic. In males the gubernaculum was long and thin, and together with the processus vaginalis it reached the inguinal region between the pubis

* Permanent address: Department of Anatomy, University of Hong Kong, 5 Sassoon Road, Hong Kong.

and the base of the scrotal anlagen (Fig. 1c, d). In females, the gubernaculum was short and thick, entering a shallow processus vaginalis which only reached as far as the abdominal transversus muscle.

Further evidence of these dimorphisms was found in nineteen karyotypically sexed embryos (7 males and 12 females) in the last 6 days of pregnancy. As early as day-5 prepartum, when the indifferent gonads were only visible as thickened ridges on the medial side of the mesonephros, scrotal or mammary anlagen were already present in males and females respectively. The gubernaculum and processus vaginalis were also sexually dimorphic in these embryos.

We conclude that sexual differentiation of the scrotum, mammary glands, gubernaculum and processus vaginalis has taken place well before gonadal differentiation in the wallaby, and therefore cannot be hormone-dependent. In support of this, androgen administration to female pouch-young of the Virginia opossum, *Didelphis virginiana*, and the tammar did not stimulate scrotal or inhibit mammary or pouch development^{4,10}. In the grey opossum, *Monodelphis domestica*, oestradiol benzoate given to neonatal males inhibited testicular development and masculinization of the reproductive tract, but did not inhibit scrotal development¹¹. This is in contrast to eutherian mammals, where scrotal development and mammary inhibition are androgen-dependent events^{12,13} following testicular differentiation.

Although the sexual differentiation of Wolffian and Mullerian duct derivatives in marsupials is under the control of testicular hormones^{4,10}, the scrotum, gubernaculum, processus vaginalis, mammary gland and pouch are not derived from these two duct systems, and their differentiation is hormone-independent. There are also some examples of apparent sex differences which

precede gonadal differentiation in eutherian mammals. Blastocoele formation is faster and somite number is greater in male mouse embryos^{14,15}, and rat males are heavier and contain more protein on day 12 of gestation¹⁶.

In conclusion, the conventional view that all somatic sexual dimorphisms in mammals are a consequence of gonadal hormone secretion can no longer be supported. Some sexual dimorphisms definitely precede gonadal differentiation in marsupials and eutherians, suggesting earlier effects of hitherto unrecognized sex-linked genes.

We thank Dr T. P. Fletcher for assistance with the animals. This study was supported by Australian Research Grants Scheme and Monash University. Animals were held under permit from the Victorian Department of Conservation, Forests and Lands.

Received 23 November 1987; accepted 4 January 1988.

- Ohno, S. *Sex Chromosomes and Sex-linked Genes* (Springer, Berlin, 1967).
- Jost, A. *Archs Anat. microsc. Morph. exp.* **36**, 271-315 (1947).
- Jost, A., Vigier, B., Prepin, J. & Perchellet, J. P. *Recent Prog. Horm. Res.* **29**, 1-35 (1973).
- Tyndale-Biscoe, C. H. & Renfree, M. B. *Reproductive Physiology of Marsupials* (Cambridge University Press, 1987).
- Evans, E. P., Burtenshaw, M. D. & Ford, C. E. *Stain Technol.* **47**, 229-234 (1972).
- Aherne, W. A. & Dunnill, M. S. *Morphometry* (Arnold, London, 1982).
- Picon, R. *Archs Anat. microsc. Morph. exp.* **74**, 81-86 (1985).
- Raeside, J. I. *Archs Anat. microsc. Morph. exp.* **74**, 87-90 (1985).
- Pelliniemi, L. J. *Archs Anat. microsc. Morph. exp.* **74**, 76-80 (1985).
- Renfree, M. B., Short, R. V. & Shaw, G. in *Genetic Markers of Sex Differentiation* (ed. Haseltine, F.) 1-15 (Plenum, New York, 1987).
- Fadem, B. H. & Tesoriero, J. V. *Biol. Reprod.* **34**, 711-776 (1986).
- Wilson, J. D., George, F. W. & Griffin, J. E. *Science* **211**, 1278-1284 (1981).
- Kratochwil, K. *Dev. Biol.* **61**, 358-365 (1977).
- Tsunoda, Y., Tokunaga, T. & Sugie, T. *Gamete Res.* **12**, 301-304 (1985).
- Sellar, M. J. & Perkin-Cole, K. J. *J. Reprod. Fert.* **79**, 159-161 (1987).
- Scott, W. J. & Holsen, J. F. *J. Embryol. exp. Morph.* **40**, 259-263 (1977).

Deletions of muscle mitochondrial DNA in patients with mitochondrial myopathies

I. J. Holt, A. E. Harding & J. A. Morgan-Hughes

Department of Clinical Neurology, Institute of Neurology, Queen Square, London WC1N 3BG, UK

In vitro studies of muscle mitochondrial metabolism in patients with mitochondrial myopathy have identified a variety of functional defects of the mitochondrial respiratory chain, predominantly affecting complex I (NADH-CoQ reductase) or complex III (ubiquinol-cytochrome c reductase) in adult cases¹⁻³. These two enzymes consist of ~36 subunits, eight of which are encoded by mitochondrial DNA (mtDNA)⁴⁻⁶. The increased incidence of maternal, as opposed to paternal, transmission in familial mitochondrial myopathy suggests that these disorders may be caused by mutations of mtDNA^{7,8}. Multiple restriction endonuclease analysis of leukocyte mtDNA from patients with the disease, and their relatives, showed no differences in cleavage patterns between affected and unaffected individuals in any single maternal line. When muscle mtDNA was studied, nine of 25 patients were found to have two populations of muscle mtDNA, one of which had deletions of up to 7 kilobases in length. These observations demonstrate that mtDNA heteroplasmy can occur in man and that human disease may be associated with defects of the mitochondrial genome.

The term mitochondrial myopathy (MM) describes a heterogeneous group of disorders which share the common feature of major structural mitochondrial abnormalities in skeletal muscle. The clinical features are variable and include proximal muscle weakness, external ophthalmoplegia, pigmentary retinopathy, dementia, ataxia, seizures, movement disorders, stroke-like episodes, deafness, peripheral neuropathy, and

cardiac conduction defects in various combinations, developing in childhood or adult life^{3,9}. Twenty per cent of patients with MM have similarly affected relatives, and the ratio of maternal to paternal transmission is ~9:1^{8,10}.

To test the hypothesis that MM may be caused by mutations of mtDNA⁸, we studied leukocyte mtDNA from 38 patients with MM, 44 of their unaffected relatives, and 35 control subjects, by means of digestion with 28 restriction endonucleases (REs) and hybridization of the fragments to ³²P-labelled HeLa-cell mtDNA. Previously unreported mtDNA polymorphisms were identified in MM families and controls, but no differences in restriction-fragment patterns were observed between affected and unaffected individuals in the same maternal line, and there was no evidence of major deletion or insertion of mtDNA in patients¹¹.

Muscle mtDNA was analysed in 25 cases of MM which had been studied biochemically¹, and 18 control subjects who were either normal or had unrelated neurological disorders. In nine patients with MM, digestion of muscle mtDNA with REs cleaving the mitochondrial genome at one or two sites¹² (for example, *Bam*HI, *Bgl*II, *Bst*EII, *Pst*I, *Pvu*II, *Kpn*I and *Xho*I) resulted in two large fragments which hybridized to labelled HeLa-cell whole mtDNA, whereas only one was observed in mtDNA from blood (Fig. 1). No extra mtDNA fragments were observed in muscle from control subjects.

We concluded from these findings that muscle from these nine patients contained two populations of mtDNA, one of which was partly deleted. The deletions were mapped by further RE analysis and hybridization with 14 small fragments of HeLa-cell mtDNA (Fig. 2). The location and extent of the deletions observed in two patients are shown in Fig. 3. In case 1, a male with complex I deficiency, the abnormal mtDNA was ~11.6 kilobases (kb) in length compared to the normal band of 16.5 kb. The smaller fragment did not hybridize to probes spanning base-pair positions 2,578-5,274, and only weakly to probe 2 (positions 41-2,578) (Fig. 2). The deletion extended at least

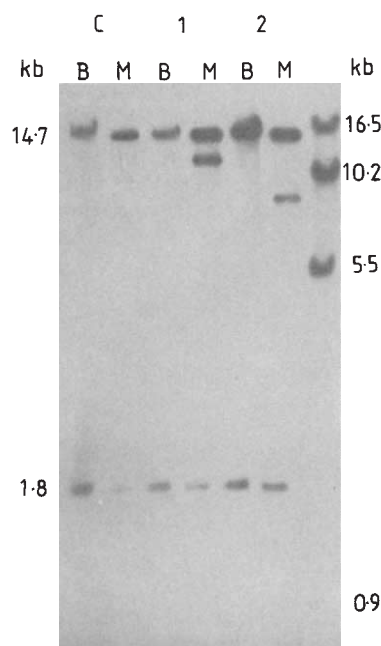


Fig. 1 Comparison between blood (B) and muscle (M) mtDNA restriction-fragment patterns in two patients with MM (1 and 2) and one normal control subject (C). After digestion with *Bst*EII (which cleaves the mitochondrial genome at positions 4,820 and 6,618, normally producing fragments of 14.7 and 1.8 kb¹²), only one large fragment (14.7 kb) is seen in mtDNA from blood in the patients, and both blood and muscle from the control. In cases 1 and 2, an extra large band (~11.6 and 8.8 kb, respectively) is seen in muscle. The 1.8 kb band is relatively faint in muscle from case 1; this fragment maps to the region which has been deleted in the abnormal population of mtDNAs (Figs 2 and 3). The lane furthest to the right shows fragments of 16.5, 10.2, 5.5 and 0.9 kb, produced by mixing *Bam*HI and *Hind*III digests of control mtDNA¹².

Methods. DNA was isolated from blood by centrifugation with 0.32 M sucrose, 10 mM Tris HCl pH 7.5, 5 mM MgCl₂, 1% Triton, and incubation of the pellet with 1 mg proteinase K, 1% sodium dodecyl sulphate (SDS). For muscle from patients and six controls, after intact mitochondria had been isolated from 5 g muscle by differential centrifugation^{1,20}, the remaining pellet (containing 70% of mitochondria in the sample) was incubated with 4 mg proteinase K and 1% SDS; frozen muscle from 12 controls was crushed with a pestle before the latter step. DNA was purified by phenol and chloroform extraction. About 5 µg of DNA was digested with the REs indicated; the resulting fragments were separated by agarose (0.8–1.8%) gel electrophoresis and transferred to nylon filters (Hybond-N, Amersham). HeLa-cell mtDNA was oligolabelled²¹ with [³²P]CTP and hybridized to the filters under conditions recommended by the manufacturers, before autoradiography at -70 °C for 24–48 h.

from position 1,301 (*Cfo*I site loss) to position 5,274 (*Eco*RV site loss) but not beyond positions 36 or 5,789 (*Sac*I and *Taq*I sites present). The deleted region thus included all or part of the coding regions for two subunits of complex I (ND), the two ribosomal RNAs, and between five and eleven transfer RNAs (Fig. 3).

Case 2 had a complex III defect and a deletion of muscle mtDNA of about 5.9 kb. Probes covering the region between positions 7,441 and 12,570 did not hybridize to the abnormal mtDNA. The deletion extended from at least position 7,440 (*Xba*I site loss) to position 12,871 (*Eco*RV site loss) with upper limits of 7,204 to 13,259 (*Msp*I and *Hinc*II sites intact). This region includes the sequences coding for three subunits of complex I, cytochrome oxidase subunits II and III (complex IV), two subunits of mitochondrial ATPase, and eight tRNAs

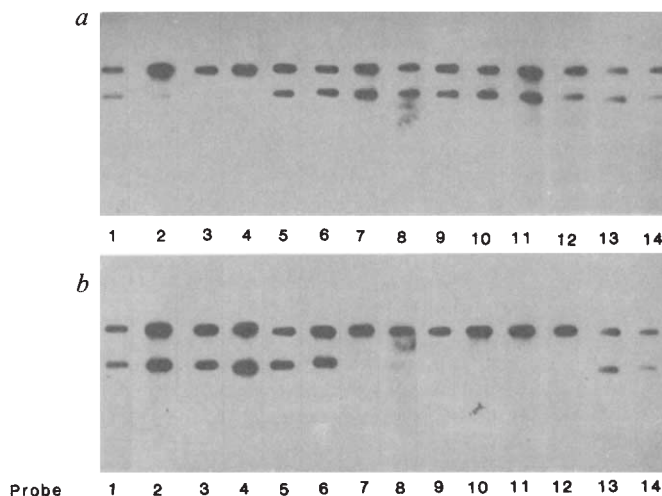


Fig. 2 Hybridization patterns of fragments of HeLa-cell mtDNA (Fig. 3) to normal and deleted populations of muscle mtDNA digested with *Bam*HI, which cleaves mtDNA at position 14,258 (ref. 12) and normally produces a single linear fragment of 16.5 kb. *a*, Case 1: hybridization to the lower fragment occurred with probes 1 and 5–14. A faint band is seen with probe 2 (base pairs 41–2,578) but the lower band did not hybridize to probes 3 and 4 (base pairs 2,578–5,274). *b*, Case 2: the lower fragment did not hybridize to probes 7–12 (base pairs 7,441–12,570) but is seen with probes 1–6, 13 and 14.

Methods. Fourteen fragments of mtDNA cloned in M13 were oligolabelled²² with [³²P]CTP and hybridized to the filters prepared as in Fig. 1.

(Fig. 3). The other seven cases of MM with two populations of mtDNA in muscle had similar, but not identical, deletions of this region. Four had complex I defects, one complex III deficiency, and in the other two, biochemical studies were normal. Densitometry suggested that the deleted mtDNAs were present in proportions ranging from 18 to 79% of total muscle mtDNA in these nine patients (37 and 54% in cases 1 and 2, respectively).

These observations have two important implications. First, they demonstrate conclusively that heteroplasmy for mtDNA occurs in man. Previous studies have suggested that there is little significant nucleotide divergence between mtDNA molecules within an individual human (or maternal line)^{13,14}, although heteroplasmy has been reported in the Holstein cow^{15,16}. A deletion of a proportion of mtDNAs in a mixture of tissues, involving a similar region to that in case 2, has recently been described in two wild mice which were probably related¹⁷.

Second, we have established that there is a molecular basis for the contention that defects of the mitochondrial genome may be associated with disease in man. In case 1, the loss of a proportion of ribosomal RNAs and two NADH-CoQ reductase (ND) subunits correlates well with the biochemical demonstration of complex I deficiency. The other patients with complex I defects also had deletions which included ND subunits. In the two cases with complex III deficiency, the relationship between the site of the deleted regions of mtDNA and that of the biochemical lesion is less clear, but the function of complex I cannot be measured polarographically in the presence of a defect further along the respiratory chain. Also, the loss of a number of tRNAs would be expected to affect the synthesis of all of the mitochondrially encoded respiratory chain subunits.

The presence of mtDNA heteroplasmy in some patients with defects of the respiratory chain provides one explanation for their variable expression and distribution in different tissues.

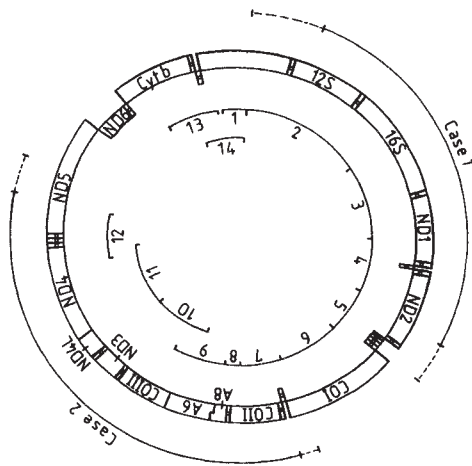


Fig. 3 Map of the human mitochondrial genome showing the location of coding regions and probes used (see Fig. 2); the extent of the deleted regions in one population of muscle mtDNA in cases 1 and 2 is also indicated (interrupted lines at each end of the two deletions define upper and lower limits; see text). Coding regions projecting inwards are on the light chain of the mtDNA molecule, the remainder on the heavy chain. ND, NADH dehydrogenase (NADH-CoQ reductase) subunits; A, ATPase; CO, cytochrome oxidase; 12S and 16S, ribosomal RNAs; | indicates tRNA.

The origin of the deletions is unclear; only one of our patients had an affected relative (his niece, who was unavailable for study). It seems probable that the deletions in the others arose during oogenesis, and that random partitioning of the two populations of mtDNA occurred during fetal development¹⁸. The survival of the deleted mtDNA molecules in muscle is compatible with the observation that the number of muscle fibres does not increase significantly after early fetal life¹⁹. Frequent cell division in leukocyte precursors could, on the other hand, select against the survival of cells containing genetically defective mitochondria.

We thank Drs G. Attardi and M. King for the HeLa-cell mtDNA probes, Drs D. Hayes and M. Cooper and Prof. J. B. Clark for performing the biochemical studies, and Marjorie Gilbert for technical assistance. This work was supported by the Muscular Dystrophy Group, the Research Trust for Metabolic Diseases of Childhood, and the Brain Research Trust.

Received 3 December 1987; accepted 21 January 1988.

1. Morgan-Hughes, J. A. *et al.* *Brain* **100**, 617-640 (1977).
2. Morgan-Hughes, J. A., Hayes, D. J., Cooper, M. & Clark, J. B. *Biochem. Soc. Trans.* **13**, 648-650 (1985).
3. Petty, R. K. H., Harding, A. E. & Morgan-Hughes, J. A. *Brain* **109**, 915-938 (1986).
4. Hatefi, Y. A. *Rev. Biochem.* **54**, 1015-69 (1985).
5. Chomyn, A. *et al.* *Nature* **314**, 592-597 (1985).
6. Chomyn, A. *et al.* in *Achievements and Perspectives of Mitochondrial Research Vol. II: Biogenesis* (eds Quagliariello, E. *et al.*) 259-275 (Elsevier, Amsterdam, 1985).
7. Hodgson, P., Bradley, W. G. & Jenkinson, M. J. *neuro. Sci.* **16**, 343-370 (1972).
8. Egger, J. & Wilson, J. N. *Eng. J. Med.* **309**, 142-145 (1983).
9. DiMauro, S., Bonilla, E., Zeviani, M., Nakagawa, M. & DeVivo, D. C. *Ann. Neurol.* **17**, 521-538 (1985).
10. Harding, A. E., Petty, R. K. H. & Morgan-Hughes, J. A. *Muscle Nerve Suppl.* **5**, 179 (1986).
11. Harding, A. E., Holt, I. J. & Morgan-Hughes, J. A. *J. med. Genet.* **4**, 243 (1987).
12. Anderson, S. *et al.* *Nature* **290**, 457-465 (1981).
13. Monnat, R. J. & Loeb, L. A. *Proc. natn. Acad. Sci. U.S.A.* **82**, 2895-2899 (1985).
14. Cann, R. L., Stoneking, M. & Wilson, A. C. *Nature* **325**, 31-36 (1987).
15. Hauswirth, W. W. & Laipis, P. J. *Proc. natn. Acad. Sci. U.S.A.* **79**, 4686-4690 (1982).
16. Hauswirth, W. W., van de Walle, M. J., Laipis, P. J. & Olivo, P. D. *Cell* **37**, 1001-1007 (1984).
17. Boursot, P., Yonekawa, H. & Bonhomme, F. *Molec. Biol. Evol.* **4**, 46-55 (1987).
18. Birky, C. W. *Ann. Rev. Genet.* **12**, 471-512 (1978).
19. Stickland, N. C. *J. Anat.* **132**, 557-579 (1981).
20. Hayes, D. J. *et al.* *J. neurol. Sci.* **71**, 105-118 (1985).
21. Feinberg, A. P. & Vogelstein, B. *Anal. Biochem.* **132**, 6-13 (1983).

Cytotoxic T-cell clones discriminate between A- and B-type Epstein-Barr virus transformants

D. J. Moss*, I. S. Misko†, S. R. Burrows*, K. Burman†, R. McCarthy* & T. B. Sculley*

* Queensland Institute of Medical Research, Bramston Tce., Herston, Brisbane 4006, Australia

† Sir Albert Sakzewski Virus Research Laboratory, Royal Children's Hospital, Herston, Brisbane 4029, Australia

Epstein-Barr virus (EBV) is the aetiological agent of infectious mononucleosis and is associated with Burkitt's lymphoma and nasopharyngeal carcinoma¹. The virus is harboured for life in all previously infected individuals² and is apparently controlled by a population of EBV-specific memory T lymphocytes, specifically activated to recognize the functionally defined lymphocyte-detected membrane antigen³. Two types (A and B) of EBV have been identified that show DNA sequence divergence within the *Bam*H1 WYH region of the genome encoding the transformation-associated antigen, Epstein-Barr nuclear antigen 2 (EBNA 2) (ref. 4). To define the function of EBNA 2 in T-cell recognition, we have compared the ability of EBV-specific cytotoxic T-cell clones to distinguish between autologous B lymphocytes transformed by A- or B-type virus. We have now isolated both CD4 and CD8 cytotoxic T-cell clones that recognize autologous A-type but not B-type transformed lymphoblastoid cell lines, thus providing the first evidence that EBV-specific T-cell recognition can be mediated by EBNA 2. As this antigen is not expressed in Burkitt's lymphoma⁵, this finding explains the failure of EBV-specific T-cell surveillance to eliminate the tumour.

To establish A- and B-type transformed lymphoblastoid cell lines (LCLs), B lymphocytes from two healthy seropositive donors (JS and DM; shown previously to be infected with the A-type EBV) were infected with either A- or B-type isolates (see legend to Fig. 1) and seeded in threefold dilutions from 10^5 cells to 10^2 cells per well. Cell lines were established by subculturing the cells seeded at the lowest concentration at which transformation was visible after 4 weeks. This procedure ensured that the established cell lines arose by infection with exogenous virus, rather than by outgrowth of endogenous EBV-infected B lymphocytes known to be present in all seropositive individuals. The virus type in each of the cell lines was checked by Western blotting and Southern analysis of viral DNA, and in all cases the transformed cell lines were found to be of the same type as the exogenous virus (Fig. 1). Histocompatibility locus antigen (HLA) typing by standard serological techniques confirmed that A- and B-type LCLs expressed class I and class II HLA antigens identical to those of the lymphocyte donor.

Lymphocytes from JS and DM were co-cultivated with autologous irradiated A-type LCL for 3 days, and the activated T cells cloned in agar were then expanded in the presence of IL-2 and stimulating cells (see Fig. 2a). T-cell phenotypes were established and cytotoxicity for EBV antigens assessed by conventional ⁵¹Cr-release assay⁶. Eight EBV-specific T-cell clones from JS and two from DM recognized the autologous A-type LCL (transformed with virus from B 95-8), but not the corresponding prototype B-type LCL (Ag876). Two typical A-type-specific clones (one from each of the two donors) are shown. The reactivity of one clone from donor JS (clone 6, CD8 phenotype) is illustrated in Fig. 2a. This clone consistently (in six experiments) showed a high degree of lysis (from 70% to 95%) towards a panel of four A-type LCLs, but an insignificant level (0.2% to 2%) of lysis to four autologous B-type LCLs or cell lines from HLA-unrelated donors. Cytotoxicity of JS clone 6 towards the autologous A-type LCL was essentially inhibited by monoclonal antibody to class I, but not to class II, deter-

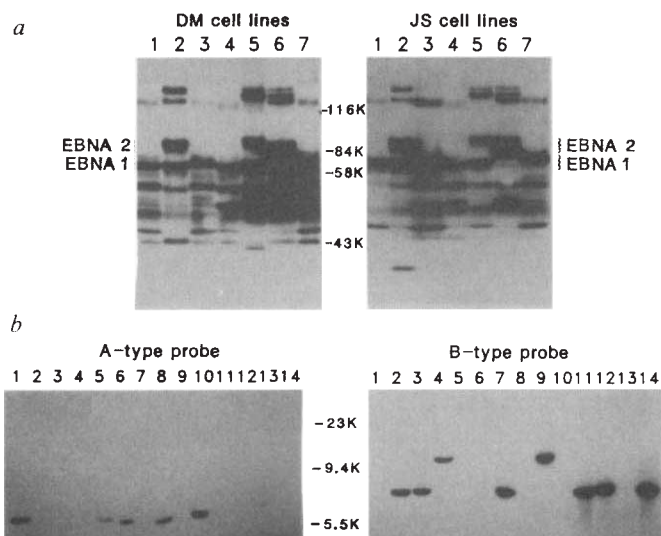


Fig. 1 Cell lines were established from two EBV seropositive donors by infecting their B lymphocytes with three different A-type isolates of EBV: B95-8 (ref. 12), IARC-BL74 (ref. 12), IARC-BL36 (unpublished data), and five different B-type isolates: Ag876 (ref. 4), BAB (unpublished data), L4 (ref. 13), QIMR-KTU (unpublished data), QIMR-JSM (unpublished data). *a*, Electrophoresis on 5–15% (w/v) polyacrylamide slab gels and immunoblotting with a serum known to react with EBNA 1 and EBNA 2 proteins were both performed as described previously¹⁴. DM and JS cells were infected with EBV obtained from the following cell lines: lane 1, QIMR-JSM; lane 2, IARC-BL74; lane 3, BAB; lane 4, L4; lane 5, B95-8; lane 6, IARC-BL36; lane 7, Ag876. *b*, EBNA 2 gene typing. DNA (5 µg) from each of the cell lines was digested by *Bam*H1, and the DNA fragments separated on 0.8% agarose gel and transferred to a nitrocellulose filter. Two parallel filters were hybridized with inserts of pM-*Bam*H2 and pJ-*HKA*7 respectively¹². DNA was isolated from DM and JS cell lines established by infection with A- and B-type virus isolates: lane 1, JS/IARC-BL74 (B cells from donor JS, transformed with virus from IARC-BL74); 2, JS/BAB; 3, JS/QIMR-JSM; 4, JS/L4; 5, JS/B95-8; 6, JS/IARC-BL36; 7, JS/Ag876; 8, DM/IARC-BL36; 9, DM/L4; 10, DM/B95-8; 11, DM/QIMR-KTU; 12, DM/QIMR-JSM; 13, DM/IARC-BL74; 14, DM/Ag876. Size markers were included in the gels and their relative molecular masses (*a*) or sizes in kilobase (kb) pairs (*b*) is indicated.

minants (Fig. 2a). The A-type specificity of this clone was confirmed in competition experiments (Fig. 4a) in which unlabelled A-type, but not B-type, targets, or an allogeneic target, successfully competed with the cytotoxicity of JS clone 6 towards the autologous A-type-transformed LCL (JS/B95-8).

The ability of clones to discriminate between LCLs infected with A- or B-type virus was not confined to T cells of the CD8 phenotype. Thus two CD4 T-cell clones were established from donor DM that recognised the autologous A-type, but failed to lyse the corresponding B-type LCL (Fig. 3a). As with the CD8 clones this result was confirmed in cold target competition experiments (Fig. 4b), but, in contrast, cytotoxicity was inhibited by antibody to class II determinants (Fig. 3a). The inhibition of CD4 and CD8 T-cell-mediated cytotoxicity by monoclonal antibodies directed against class II and class I public determinants confirms previous findings⁵.

The consistent failure to lyse the B-type LCLs with A-type-specific clones is not related to an intrinsic inability of these target cells to undergo lysis or to present antigen. Thus the B-type LCLs from JS were susceptible to lysis by allospecific T cells (Fig. 2b), whereas cross-reacting T-cell clones (for example, clone 7, CD8 phenotype) from DM showed an equal degree of lysis towards A- and B-type LCLs (Fig. 3b), and their cytotoxicity towards the A-type target was subject equally to cold target

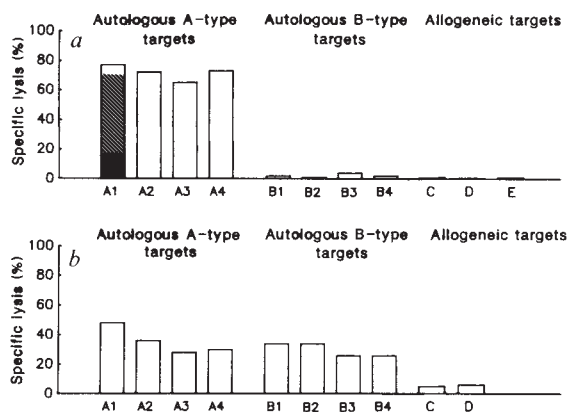


Fig. 2 *a*, Percentage specific lysis of autologous LCLs transformed with A-type (A1–A4) or B-type virus (B1–B4), and allogeneic LCLs (C, D, E) from HLA-unrelated donors, by JS clone 6. Cell lines: A1, JS/IARC-BL74; A2, JS/B95-8; A3, JS/IARC-BL36; A4, JS (spontaneous line); B1, JS/L4; B2, JS/QIMR-JSM; B3, JS/BAB; B4, JS/Ag876; C, DJM; D, IM; E, K562. *b*, Percentage specific lysis of LCLs from donor JS and allogeneic cell lines (see Fig. 2a) by allospecific effector T cells.

Methods. *a*, T-cell clones were generated as previously described⁶. Peripheral blood mononuclear cells (2×10^6 per well) were stimulated with the γ -irradiated (8,000 R) autologous A-type (JS/B95-8) LCL (10^4 per well) in multi-well plates (Costar) in RPMI 1640 medium supplemented with 10% heat-inactivated fetal calf serum (HIFCS) and incubated at 37 °C in 5% CO₂ in air. At 3 d, cells were dispersed, collected and seeded into 0.35% agarose (Sea-Plaque) containing RPMI 1640, 20% HIFCS, 25% IL-2-containing supernatant from MLA-144 T-cell cultures and recombinant IL-2 (30 U ml^{-1}). Colonies were collected after three further days and amplified in culture with bi-weekly restimulation with IL-2 and irradiated autologous LCL. T-cell markers were assayed with monoclonal antibodies OKT 3, 4 and 8 (Ortho) and fluorescein-labelled goat anti-mouse IgG antibody (Kirkegaard and Perry Lab) by flow cytometry. T-cell clones were tested for cytotoxicity in a standard 5 h ⁵¹Cr-release assay⁶ using an effector:target ratio of 3:1. Monoclonal antibodies specific for the monomorphic determinants on class I (1,620, ■) or class II (1,622, □) HLA antigens (provided by I. McKenzie) were tested for their efficacy in blocking autologous killing (final antibody dilution 1/40) in the standard 5 h ⁵¹Cr-release assay. The cell surface phenotype of JS colony 6 was 94.0% OKT 3, 4.2% OKT 4, 90% OKT 8. The HLA serotype of donor JS is A1, 2; B8, 51; DR4, w6. *b*, prepared by co-cultivating lymphocytes (2×10^6 per well) from donor DM with irradiated JS/IARC-BL74 (responder to stimulator ratio, 20 to 1) for 7 d and then restimulating for a further 4 d. The resorted T cells were used as effectors in a 5 h ⁵¹Cr-release assay at an effector to target ratio of 10 to 1.

competition by both A- and B-type LCLs (data not shown). The inclusion in the target panel of autologous LCLs grown in fetal calf serum or in human serum, together with their equal levels of lysis by the A-type-specific clone (Fig. 3a) and cross-reactive clone (Fig. 3b), showed that the specificity patterns of lysis were independent of fetal calf serum-specific effect⁷.

These experiments show the isolation for the first time of both CD4 and CD8 clones of EBV-specific cytotoxic T cells that discriminate between cell lines infected with the two types of EBV. The chief DNA-sequence differences between the two virus types are within the BYRF1 reading frame, whose product is EBNA 2, though some minor differences occur within another virus-coded nuclear protein, the leader protein^{4,8}. Therefore, it is likely that regions within one or both of these nuclear proteins contain the relevant epitope for EBV-specific T-cell recognition for the clones described. This approach provides us with the opportunity of defining lymphocyte-detected membrane antigen epitopes in a way similar to that used in human influenza⁹, where the major target epitopes are associated with a nuclear

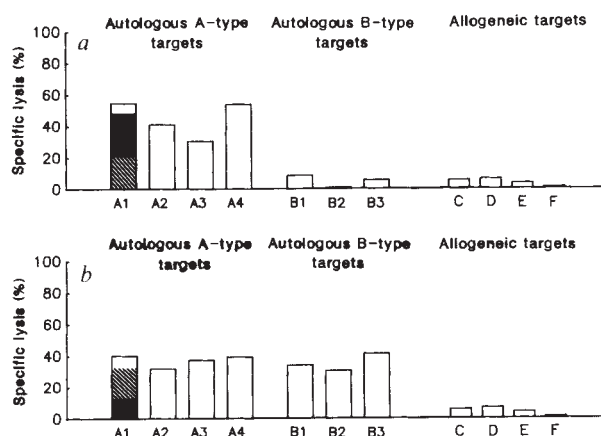


Fig. 3 *a*, Percentage specific lysis of autologous LCLs transformed with A-type (A1-A4) or B-type (B1-B3) EBV, and allogeneic LCLs (C-F) from HLA-unrelated donors by A-type-specific DM clone 4. Cell lines: A1, DM/B95-8 (grown in fetal calf serum); A2, DM/B95-8 (grown in human serum); A3, DM/IARC-BL74; A4, DM/IARC-BL36; B1, DM/Ag876; B2, DM/QIMR-KTU; B3, DM/QIMR-JSM; C, LD; D, DJ; E, SH; F, K562. T-cell clones were generated by autologous LCL (DM/B95-8) stimulation and assayed for phenotypic markers, cytotoxicity and monoclonal antibody inhibition as outlined in Fig. 2*a* legend. The cell surface phenotype of DM clone 4 was 91.3% OKT3, 91.3% OKT4, 2.6% OKT8. The HLA serotype of donor DM is A24, 29; B12, 47; DR1, 7. Monoclonal antibodies against class I (1,620 ■) and class II (1,622, ▨) determinants were used as in Fig. 2*a*. *b*, Percentage specific lysis of autologous and allogeneic LCL (targets as in legend to Fig. 3*a*) by cross-reactive EBV-specific DM clone 7 generated in culture at a responder to stimulator ratio of 2 to 1. The cell surface phenotype of DM clone 7 was 95.9% OKT3, 3.0% OKT4, 91.5% OKT8. Inhibition of cytotoxicity by monoclonal antibodies as in Fig. 2*a* legend.

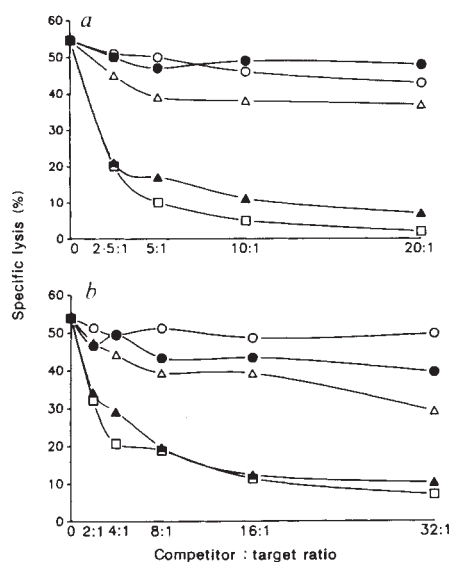


Fig. 4 *a*, Cold-target inhibition of specific lysis of autologous LCL (JS/B95-8) target cells by JS clone 6. The curves represent autologous lysis recorded with unlabelled autologous A-type competitor cells (JS/IARC-BL74, □; JS/IARC-BL36, ▲) unlabelled autologous B-type competitor cells (JS/QIMR-JSM, △; JS/L4, ○) or allogeneic LCL from an HLA-unrelated donor DM (●). Effector cell: labelled target cell ratio 3:1. *b*, Cold target inhibition of specific lysis of autologous LCL (DM/B95-8) target cells by DM clone 4 in the presence of unlabelled A-type autologous competitor cells (DM/B95-8, □; DM/QIMR-BL74, ▲), unlabelled B-type competitor cells (DM/Ag876, △; DM/QIMR-KTU, ○) or allogeneic LCL from an HLA-unrelated donor SH (●). Effector cell: labelled target cell ratio 5:1.

protein. Evidence exists that the EBV latent membrane protein can act as a target for EBV-specific cytotoxic T cells¹⁰. This observation is relevant because non-nuclear proteins have also been implicated as targets for T-cell recognition in the human influenza system. The present data in no way excludes the involvement of non-nuclear proteins in EBV-specific T-cell recognition. Indeed, it is possible that the cross-reacting clones described here recognize polypeptides encoded by sequences outside the *Bam*H1 WYH genome region.

The present study has important implications for the biology of the EBV, particularly the ability of EBV-associated tumours to proliferate in Burkitt's lymphoma patients, most of whom appear to have no EBV-specific T-cell dysfunction¹¹. It is possible that EBV-specific T-cell immunosurveillance, which is extremely efficient at controlling the outgrowth of virus-infected normal B cells³, may not recognize fresh tumour cells from these patients because EBNA 2 is not expressed in these cells.

This work was supported by grants from the National Health and Medical Research Council of Australia and the Queensland Cancer Fund. We thank Drs G. M. Lenoir and A. B. Rickinson for providing cell lines and N. Strahan for the HLA typing of cell lines.

Received 29 September; accepted 23 December 1987.

- Epstein, M. A. & Achong, B. G. in *The Epstein-Barr Virus: Recent Advances* (ed. Epstein, M. A. & Achong, B. G.) 1-11 (Heinemann, London, 1986).
- Nilsson, K., Klein, G., Henle, W. & Henle, G. *Int. J. Cancer* **8**, 443-450 (1971).
- Rickinson, A. B. *et al. Cancer Res.* **41**, 4216-4221 (1981).
- Dambaugh, T., Hennessy, K., Chamnankit, L. & Kieff, E. *Proc. natn. Acad. Sci. U.S.A.* **81**, 7632-7636 (1984).
- Rowe, M. *et al. EMBO J.* **6**, 2743-2751 (1987).
- Misko, I. S., Pope, J. H., Hutter, R., Soszynski, T. D. & Kane, R. G. *Int. J. Cancer* **33**, 239-243 (1984).
- Misko, I. S., Soszynski, T. D., Kane, R. G. & Pope, J. H. *Clin. Immun. Immunopath.* **32**, 285-297 (1984).
- Addinger, H. K., Delius, H., Freese, U. K., Clarke, J. & Bornkamm, G. W. *Virology* **141**, 221-234 (1985).
- Townsend, A. R. M. *et al. Cell* **44**, 959-968 (1986).
- Thorley-Lawson, D. A. & Israelsohn, E. S. *Proc. natn. Acad. Sci. U.S.A.* **84**, 5384-5388 (1987).
- Rooney, C. M., Rickinson, A. B., Moss, D. J., Lenoir, G. M. & Epstein, M. A. *IARC scient. Publs.* **60**, 249-264 (1985).
- Zimmer, U. A. *et al. Virology* **154**, 56-66 (1986).
- Young, L. *et al. J. gen. Virol.* (in the press).
- Sculley, T. B., Walker, P. J., Moss, D. J. & Pope, J. H. *J. Virol.* **52**, 88-93 (1984).

Homologies between gap junction proteins in lens, heart and liver

Joerg Kistler*, David Christie† & Stanley Bullivant*

* Centre for Gene Technology, Department of Cellular and Molecular Biology and † Department of Biochemistry, University of Auckland, Auckland, New Zealand

The cells in the mammalian lens are electrically¹ and metabolically² coupled with each other by a network of gap junctions. These are clusters of transmembrane channels by which the fibre cells situated deeper in the lens communicate through the epithelium with the aqueous humour, the source of nutrients for the lens. Hence gap junctions are important for lens transparency. The gap junction proteins in the mammalian lens have not yet been identified with certainty. A putative fibre gap junction protein of relative molecular mass 26,000 (26K)³ is not related to those from other tissues, such as the liver 28K junction component^{4,5}. Another lens membrane protein with *M_r* 70K (MP70) has also been localized in the lens fibre gap junctions^{6,7}. Here we demonstrate by amino-terminal sequence analysis that MP70 and its *in vivo*-processed form, MP38 (ref. 8), belong to a wider family of gap junction proteins⁹. With this new data on the lens, homologies between gap junction proteins now extend to organs derived from all three embryonic layers, endoderm (liver), mesoderm (heart) and ectoderm (lens).

Fig. 1 Visualization of fibre gap junctions by anti-MP70 immunofluorescence microscopy. Each fluorescent plaque represents a gap junction structure based on previous correlations with immunoelectron microscopy⁷.

Methods. A lens of a less-than-1-year-old sheep has been cryosectioned in tangential direction through the outermost equatorial region. Formaldehyde-fixed sections

were labelled with anti-MP70 antibodies produced by hybridoma 6-4-B2-C6 (ref. 6), and bound antibodies visualized with secondary goat-anti-mouse IgM conjugated to fluorescein. Scale bar 10 μ m.

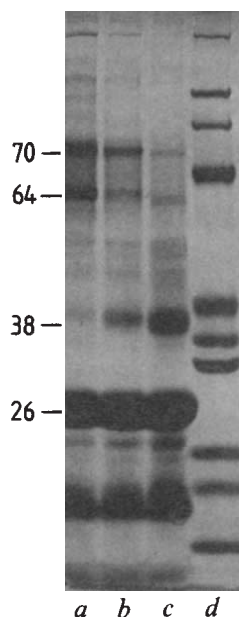
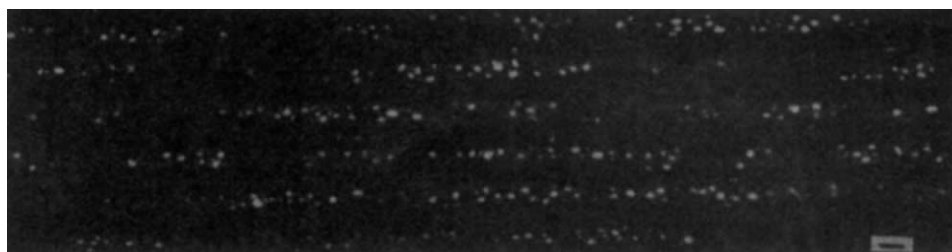


Fig. 2 Stained gels of membranes isolated from *a*, outer, *b*, intermediate and *c*, inner lens region. MP70 and MP64 generally occur together, the latter being a post-mortem breakdown product of MP70. MP38 is derived from MP70 by limited proteolysis *in vivo*⁸. *d*, Size markers from top, $M_r \times 1,000$: 205, myosin; 116, β -galactosidase; 97, phosphorylase *b*; 66, bovine serum albumin; 45, ovalbumin; 36, glyceraldehyde-3-phosphate dehydrogenase; 29, carbonic anhydrase; 24, trypsinogen; 20, soybean trypsin inhibitor; 14, α -lactalbumin.

Methods. Less-than-1-year-old sheep lenses were dissected and the membranes isolated by alkaline stripping as described previously⁸. Membranes (100 μ g protein per lane) were solubilized without heating in sample buffer immediately before loading the gels. SDS-PAGE (10%) was according to Laemmli¹⁵.

Using anti-MP70 immunofluorescence microscopy of sheep lens sections, fibre gap junctions are visible as fluorescent plaques in otherwise unlabelled plasma membranes (Fig. 1). The distribution of junctional plaques varies with the developmental state of the fibre cells, and in the older cells deeper in the lens, these junctions fragment to smaller sizes as part of an ageing process¹⁰. Along with this fragmentation, MP70 is converted to MP38 by limited proteolysis⁸. When proteins of membranes isolated exclusively from the outer lens region are examined on stained gels, these show predominantly MP70 (and its post-mortem degradation product MP64) and relatively little MP38 (Fig. 2*a*). The situation is reversed in the membranes isolated from the inner lens region, which contain mostly the processed form, MP38 (Fig. 2*c*).

For further characterization, we analysed the amino-terminal sequence in samples of MP70, MP64 and MP38 electro-blotted on to Immobilon PVDF transfer membranes¹¹. The sequences were identical for all three polypeptides, confirming the precursor-product relationships (Fig. 3). This result also means that the *in vivo*-processed form MP38 is derived from MP70 by cleavage at a site in the carboxy-terminal half of the molecule, leaving the amino-terminal intact. As this cleavage occurs in intact cells, the cleavage site is probably located on the cytoplasmic side of the fibre plasma membrane. In all cases, position 8 was Asn and Arg, position 9 was both Ile and Leu, and in the case of MP38, both Ala and Val were found in position 13. We have repeatedly obtained these doublets for otherwise unambiguous sequences, which indicates some polymorphism of MP70 and its cleavage products.

The longest amino-terminal sequence was determined for MP38. We compared this sequence with the previously published sequences from the gap junction proteins in rat heart and liver⁹ and found extensive homologies. Using as a criterion the conservation of residues in all four gap junction proteins, positions Trp-3, Leu-6, Leu-10, Val-13, Ser-17 and Thr-18 appear important. Furthermore, the amino-terminal sequence for MP70,

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
LENS 70	Gly	Asp	Trp	(?)	Phe	Leu	Gly	Asn Arg	Ile Leu	Leu										
LENS 64	Gly	Asp	(?)	Ser	Phe	Leu	Gly	Asn Arg	Ile Leu	Leu	Glu	Asn								
LENS 38	Gly	Asp	Trp	Ser	Phe	Leu	Gly	Asn Arg	Ile Leu	Leu	Glu	Asn	Ala Val	Gln	Glu	His	Ser	Thr	Val	Ile
HEART 47	Ala	Asp	Trp	Ser	Ala	Leu	Gly	Lys	Leu	Leu	Asp	Lys	Val	Gln	Ala	Tyr	Ser	Thr	Ala	Gly
LIVER 28	Met	Asn	Trp	Thr	Gly	Leu	Tyr	Thr	Leu	Leu	Ser	Gly	Val	Asn	Arg	His	Ser	Thr	Ala	Ile
LIVER 21	(?)	Asp	Trp	Gly	Thr	Leu	Gln	Ser	Ile	Leu	Gly	Gly	Val	Asn	Lys	His	Ser	Thr		

Fig. 3 The amino-terminal residues of MP70/64 and MP38 in comparison with those previously determined for the 47K component from heart and the 28/21K components from liver gap junctions^{9,16}. Sequences are identical for the lens proteins. Amino-terminal sequences are 40% identical for lens and liver, and 50% identical for lens and heart.

MP64 and MP38 is identical to that derived from a gap junction-related complementary DNA prepared from lens¹². But we found no homology between these gap junction proteins and the 26K, putative fibre gap junction protein for which the complete sequence is known⁴.

A number of conclusions can be drawn from these results: a divergent family of gap junction proteins exists of M_r ranging from 21K to 70K. The differences at the molecular level contrast to the more or less universal morphology of gap junctions. The divergence of gap junction proteins, however, indicates the adaptation of these structures to the specialized requirements in the organs. In the case of the liver 28K component, channel-forming properties have already been demonstrated^{13,14}. Hence MP70, either alone or with other proteins, probably forms the channels in the lens fibres. Further, with MP70 and MP38 now identified as gap junction proteins, the lens fibre gap junctions become the first reported to undergo proteolytic processing *in vivo*. Assays will now have to be developed to study what effect such processing has on the permeability of gap junctions and on the maintenance of lens transparency.

We thank Dr A. R. Bellamy for constructive comments on this manuscript. This work was supported by grants from the Medical Research Council of New Zealand.

Received 24 November 1987; accepted 7 January 1988.

1. Rae, J. L. & Mathias, R. T. in *The Ocular Lens* (ed. H. Mäkelä) 93–121 (Marcel Dekker, N.Y., 1985).
2. Goodenough, D. A., Dick, J. S. B. & Lyons, J. E. *J. Cell Biol.* **86**, 576–589 (1980).
3. Revel, J. P., Yancey, S. B., Nicholson, B. & Hoh, J. *Ciba Fdn Symp.* **125**, 108–127 (1986).
4. Gorin, M. B., Yancey, S. B., Cline, J., Revel, J. P. & Horwitz, J. *Cell* **39**, 49–58 (1984).
5. Paul, D. *J. Cell Biol.* **103**, 123–134 (1986).
6. Kistler, J., Kirkland, B. & Bullivant, S. *J. Cell Biol.* **101**, 29–35 (1985).
7. Grijters, W. T. M., Kistler, J., Bullivant, S. & Goodenough, D. A. *J. Cell Biol.* **104**, 565–572 (1987).
8. Kistler, J. & Bullivant, S. *Invest. Ophthalmol. Vis. Sci.* **28**, 1687–1692 (1987).
9. Nicholson, B. *et al. Nature* **329**, 732–734 (1987).
10. Grijters, W. T. M., Kistler, J. & Bullivant, S. *J. Cell Sci.* **88**, 351–360 (1987).
11. Matsudaira, P. *J. Biol. Chem.* **262**, 10035–10038 (1987).
12. Beyer, E. C., Paul, D. L. & Goodenough, D. A. *J. Cell Biol.* **105** (in the press).
13. Dahl, G., Miller, T., Paul, D., Voellmy, R. & Werner, R. *Science* **236**, 1290–1293 (1987).
14. Young, J. D. E., Cohn, Z. A. & Gilula, N. B. *Cell* **48**, 733–743 (1987).
15. Laemmli, U. K. *Nature* **227**, 680–685 (1970).
16. Nicholson, B., Gros, D., Kent, S., Hood, L. & Revel, J. P. *J. Biol. Chem.* **260**, 6514–6517 (1985).

Gene for a novel tRNA species that accepts L-serine and cotranslationally inserts selenocysteine

Walfred Leinfelder*, Eva Zehelein*, Marie-André Mandrand-Berthelot† & August Böck*

* Institut für Genetik und Mikrobiologie der Universität München, Maria-Ward-Strasse 1a, D-8000 München 19, FRG

† Laboratoire de Microbiologie, INSA, 20 Avenue Albert Einstein, F-69621 Villeurbanne Cedex, France

The biological requirement of the trace element selenium was recognized 40 years ago¹. Selenium is incorporated into several enzymes and transfer RNA species of both prokaryotic and eukaryotic origin². In enzymes which contain a selenopolypeptide, selenium is present as covalently bound selenocysteine which participates in the catalytic reaction³. Sequence analysis of the genes coding for two selenoproteins, formate dehydrogenase H from *Escherichia coli*⁴ and glutathione peroxidase from mouse⁵ and man⁶, demonstrated that an in-frame UGA opal nonsense codon directs the incorporation of selenocysteine. In the case of formate dehydrogenase incorporation occurs cotranslationally⁷. Recently, we identified four genes whose products are required for selenocysteine incorporation in *E. coli*⁸. We report here that one of these genes codes for a tRNA species with unique properties. It possesses an anticodon complementary to UGA and deviates in several positions from sequences, until now, considered invariant in all tRNA species^{9,10}. This tRNA is aminoacylated with L-serine by the seryl-tRNA ligase which also charges cognate tRNA_{Ser}. Selenocysteine, therefore, is synthesized from a serine residue bound to a natural suppressor tRNA which recognizes UGA.

E. coli possesses two distinct selenopolypeptides which are constituent subunits of formate dehydrogenase H (FDH_H) and formate dehydrogenase N (FDH_N)¹¹. Mutations in three genes (*selA*, *selB* and *selC*) block incorporation of selenium into these proteins but not into tRNA⁸. The genes *selA* and *selB* have recently been cloned⁸ and have been shown to code for polypeptides of relative molecular mass (M_r) 50,000 (50K) and 70K, respectively.

The unique cotranslational mechanism of selenocysteine insertion into the selenopolypeptide of FDH_H requires the existence of a tRNA species which is capable of recognizing the in-frame opal (UGA) codon within the FDH messenger RNA. We have, therefore, also cloned *selC* and determined its nucleotide

sequence to see whether it codes for such a tRNA molecule (Fig. 1). A cosmid gene bank of *E. coli* MC4100 chromosomal DNA⁸ was screened for complementation of the *selC* mutation present in *E. coli* MB07 (ref. 8) using the benzylviologen dye overlay technique⁸. The 35 kilobase (kb) DNA insert of a positive clone obtained was shortened by subcloning down to a size of 233 base pairs (bp) which still could complement the *selC* defect of MB07 (Fig. 1a).

The nucleotide sequence of the insert (Fig. 1b) has no open reading frame which could encode a potential protein product. The insert, however, contains the coding information for a putative tRNA molecule. A model of its secondary structure is shown in Fig. 2.

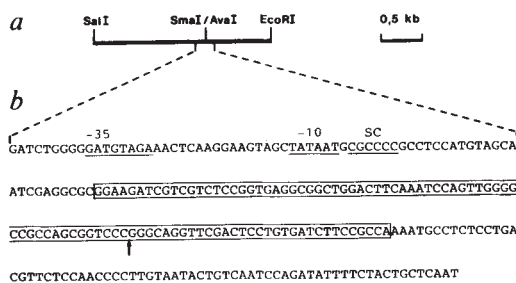
There are a number of arguments in favour of the notion that *selC* codes for a functional tRNA species which inserts selenocysteine into protein. First, plasmid pMN73 (see legend to Fig. 1) restores the capacity of MB07 (*selC*) to incorporate selenocysteine into the selenopolypeptide subunits of both formate dehydrogenases (Fig. 1c) and for formation of active FDH_H and FDH_N enzymes (data not shown). Second, the tRNA coding region of the insert of pMN73 is preceded by a –10 and –35 *E. coli* promoter sequence and, in addition, by the primary

Table 1 Charging specificity of tRNA from cells carrying *selC* on a multicopy plasmid (pMN81)

Enzyme source	Amino acid	pmoles amino acid per 100 µg tRNA	
		FM420/pUC18	FM420/pMN81
MC4100	Serine	33	217
MC4100	Leucine	144	117
MC4100	Phenylalanine	72	90
KL235, untreated	Serine	24	200
KL235, 20 min 45 °C	Serine	0.9	2
KL235, untreated	Leucine	121	97
KL235, 20 min 45 °C	Leucine	119	96
KL235, untreated	Phenylalanine	72	76
KL235, 20 min 45 °C	Phenylalanine	65	67

Bulk tRNA prepared from strain FM420 (ref. 4) carrying either vector pUC18 or plasmid pMN81 was charged to completion²³ in a standard aminoacylation assay using protein (100 µg) of a 100,000 g supernatant from MC4100 or from the *serS* mutant KL235 (ref. 18). The seryl-tRNA ligase from this strain is inactivated by incubation at 45 °C for 20 min¹⁸. The following amino acids were also tested and were found to deliver the same results as leucine and phenylalanine: cysteine, glycine, threonine, alanine, glutamate, aspartate and tryptophan.

Fig. 1 Cloning and nucleotide sequence of the *selC* gene. *E. coli* MC4100 (ref. 8) chromosomal DNA was restricted with endonuclease *Sau*3A and the fragments were cloned into cosmid vector pHC79 (ref. 16). One positive clone was obtained which complemented the *selC* lesion of strain MB07 (ref. 8). Its DNA insert (about 35 kb) was subjected to a partial digestion with *Sau*3A; the fragments were ligated into plasmid pACYC184 (ref. 16), linearized with *Bam*HI and treated with calf intestinal alkaline phosphatase. One plasmid (pMN60) able to complement the *selC* mutation of MB07 (insert size 850 base pairs (bp)) was used for the construction of further subclones: (1) pMN73 was obtained by cloning an *Ssp*I-*Nhe*I fragment into *Hinc*II-EcoRV digested pACYC184; (2) pMN81 carries the same fragment cloned into the *Sma*I site of pUC18 (ref. 16). Both pMN73 and pMN81 were positive in *selC* complementation tests. pMN81 was used for nucleotide sequence analysis of the insert with the chemical cleavage method^{12,13} using the *Sal*I site of the polylinker. Both strands were sequenced^{12,13}. The clone pMN81 (and pMN73) carries an insert of 379 bp, 233 bp are derived from MC4100 chromosomal DNA, 146 bp from the *tet* gene of pACYC184. **a** Shows the relative position of the 233-bp segment of pMN81 derived from MC4100 within an *Sal*I-EcoRI restriction fragment as determined by hybridization with fragmented chromosomal DNA. **b** Shows the nucleotide sequence of the non-coding DNA strand. The segment coding for tRNA is boxed; sequence motifs in the 5'-flanking region corresponding to the -35 and -10 promoter consensus and to the stringent control motif (SC)¹⁴ are indicated. The arrow indicates the position of the *Sma*I/*Ava*I restriction site. **c** Shows an autoradiograph of an SDS polyacrylamide gel from lysates of cells labelled with [⁷⁵Se]selenite. Lanes: 1 and 2, *E. coli* EMG29 (ref. 8) (wild-type); 3 and 4, MB07 (ref. 8) (*selC*) carrying pACYC184; 5 and 6, MB07 (*selC*) carrying pMN73. Cells were grown anaerobically in the absence (lanes 1, 3, 5) or presence (lanes 2, 4, 6) of nitrate which is the inducer of FDH_N synthesis. The positions of the 70K and 110K selenopolypeptides of FDH_H and FDH_N are indicated.



structure motif (SC in Fig. 1b) implicated in stringent control of synthesis of stable RNA in *E. coli*¹⁴. The similarity of the sequence of this 5' flanking region with those preceding other tRNA genes¹⁴ suggests that this tRNA gene is expressed.

Additional evidence comes from gene disruption experiments. The coding region of pMN73 (Fig. 1b) contains an *Sma*I restriction site. The plasmid pMN73 was opened at this *Sma*I site and the omega element¹⁵ was inserted. In a parallel experiment pMN73 was cut with *Ava*I (Fig. 1b), the recessed ends were filled in with DNA polymerase (Klenow fragment) and religated. Both treatments abolished the capacity of pMN73 to complement the *selC* lesion present in MB07.

As a final proof we localized the mutation within the *selC* gene of mutant MB07. We found that the sequence of the MB07 *selC* gene differed in a single position from the wild-type sequence: G 71 within the universally conserved tRNA sequence GTTC^{9,10} is changed to an A (Fig. 2). These results indicate that the *selC* gene codes for a tRNA species that incorporates selenocysteine into protein.

The secondary-structure model of this tRNA reveals a number of unique features absent in other *E. coli* tRNAs (Fig. 2). First, it possesses an anticodon sequence directly matching the UGA codon. The structural features of the anticodon arm correspond almost exactly with those characteristic of tRNA species which have an A at the third position in the anticodon¹⁷. Second, with a length of 95 nucleotides (derived from the DNA sequence) it is one of the largest tRNA molecules yet described^{9,10}. This is mainly due to the large size of the extra arm. Third, the amino-acid acceptor stem consists of eight base pairs instead of the seven base pairs normally found. The existence of the base pair between nucleotides 1 and 91, however, is predicted only from the gene sequence. If G 1 is eliminated by processing then the 3'-terminal -CCA is separated from the helical region of the amino-acid acceptor stem by two unpaired nucleotides which would also be unique. Finally, several nucleotide positions considered invariant in all tRNA species from *E. coli*^{9,10} deviate from the consensus. They are all clustered around the D loop in the secondary-structure model. Specifically, these changes are a G in position 9 instead of a U, two pyrimidines instead of two purines in positions 15 and 16, and a purine-pyrimidine base pair between the bases at positions 12 and 25 instead of a pyrimidine-purine pair. The latter trait is also characteristic of methionine-initiator tRNA species^{9,10}.

We performed charging experiments to determine the amino-acid specificity of the *selC* gene product. The DNA insert present in plasmid pMN73 was cloned into pUC18 to increase the gene dosage of *selC*. Bulk tRNA was then prepared from cells containing the resulting plasmid (pMN81) and from control cells harbouring pUC18 without the insert¹⁶. Table 1 shows that tRNA from cells containing pMN81 exhibits a significantly increased serine-acceptor capacity; none of the other nine amino acids tested displayed such a difference between pMN81 and pUC18 alone. Also selenocysteine could not quench this increased charging by L-serine, indicating that this tRNA cannot be charged with free selenocysteine (data not shown).

We also tested aminocylation with extracts from a mutant of *E. coli* possessing a temperature-sensitive seryl-tRNA ligase¹⁸. Heat inactivation of the enzyme abolished charging of total serine-acceptor activity of tRNA from cells carrying pUC18 and pMN81 (Table 1). This result strongly suggests that the *selC* gene product is directly charged with L-serine and that the *serS* gene product, seryl-tRNA ligase, is the enzyme responsible. This result is in accordance with the finding that the carbon skeleton of selenocysteine is derived from L-serine in eukaryotic cells²².

The results described offer a plausible mechanism for selenocysteine synthesis and incorporation that involves charging of the *selC* gene product with L-serine, modification of the serine residue esterified to tRNA at the hydroxyl group and exchange of this group for the selenol group. Involvement of an enzyme(s) which specifically recognizes part of this tRNA molecule could provide an explanation for the extreme specificity with which the selenocysteine pathway differentiates between selenium and sulphur.

A mutant in which the genes *selA* and *selB*⁸ are deleted is unimpaired in aminoacylation of the tRNA with serine (data not shown). This indicates that involvement of the products of these genes in the incorporation of selenocysteine into FDH occurs after charging of the tRNA. Experimental substantiation of the pathway proposed above may be forthcoming from functional analyses of the *selA* and *selB* gene products.

A natural opal-suppressor tRNA which can be charged by serine and phosphorylated by a kinase to O-phosphorylseryl-tRNA has been described for eukaryotic cells¹⁹. The function of this tRNA may be to insert O-phosphorylseryl-tRNA into proteins^{20,21}. But it is equally possible that this tRNA inserts selenocysteine into eukaryotic selenoproteins, like glutathione

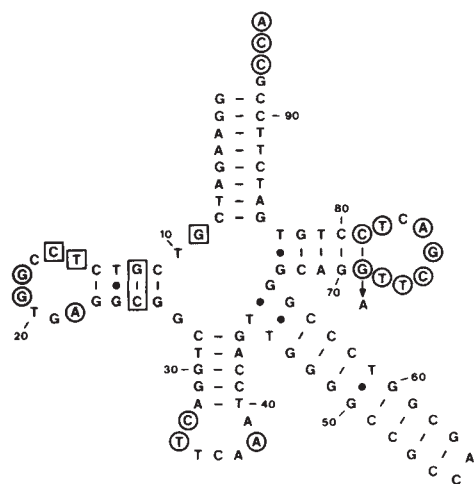


Fig. 2 Secondary-structure model of the tRNA species which complements the *selC* mutation, as derived from the DNA sequence. Bases corresponding to invariant positions^{9,10} of *E. coli* tRNA are circled, those deviating from the consensus are boxed. The point mutation in the MB07 *selC* sequence (position 71) is indicated by an arrow.

peroxidase, and that the *O*-phosphorylserine residue is simply the biochemical intermediate for selenocysteine formation.

Apart from the elucidation of the pathway of selenocysteine formation there are a number of intriguing questions which remain to be answered. The most crucial of which is the mechanism whereby the internal UGA codon is differentiated from a UGA termination codon. Context effects on the message may be involved as well as the unusual primary and secondary structural features of the selenocysteine-inserting tRNA. Another interesting question is why this tRNA specifically inserts selenocysteine and not serine or *O*-phosphorylserine into the FDH selenopolypeptide^{4,7,8}. Finally, the phylogenetic conservation of the selenocysteine insertion mechanism also prompts the speculation as to whether UGA was originally a sense codon; after introduction of oxygen into the biosphere this highly oxidizable amino acid could be maintained only in anaerobic organisms or in aerobic systems which evolved specific protection mechanisms. As a consequence, the UGA codon then could have acquired other functions such as its more familiar role in termination.

We thank Gary Sawers for many suggestions and for reading the manuscript. This work was supported by grants from the Deutsche Forschungsgemeinschaft and by the Fonds der Chemischen Industrie.

The sequence data in this publication have been submitted to the EMBL/GenBank Libraries under the accession number Y00299.

Received 23 November; accepted 7 January 1988.

1. Pinsent, J. *Biochem. J.* **57**, 10–16 (1954).
2. Stadtman, T. C. *A. Rev. Biochem.* **49**, 93–110 (1980).
3. Condell, R. A. & Tappel, A. L. *Biochim. biophys. Acta* **709**, 304–309 (1982).
4. Zinoni, F., Birkman, A., Stadtman, T. C. & Böck, A. *Proc. natn. Acad. Sci. U.S.A.* **83**, 4650–4654 (1986).
5. Chambers, I. *et al.* *EMBO J.* **5**, 1221–1227 (1986).
6. Sukenaga, Y., Ishida, K., Takeda, T. & Takagi, K. *Nucleic Acids Res.* **15**, 7178 (1987).
7. Zinoni, F., Birkman, A., Leinfelder, W. & Böck, A. *Proc. natn. Acad. Sci. U.S.A.* **84**, 3156–3160 (1987).
8. Leinfelder, W. *et al.* *J. Bact.* (in the press).
9. Sprinzl, M., Moll, J., Meissner, F. & Hartmann, T. *Nucleic Acids Res.* **13**, r1–r49 (1985).
10. Sprinzl, M., Vorderwülbecke, T. & Hartmann, T. *Nucleic Acids Res.* **13**, r51–r104 (1985).
11. Cox, J. C., Edwards, E. S. & DeMoss, J. A. *J. Bact.* **145**, 1317–1324 (1981).
12. Maxam, A. M. & Gilbert, W. *Meth. Enzym.* **65**, 499–560 (1980).
13. Gray, C. P., Sommer, R., Beck, E. & Schaller, H. *Proc. natn. Acad. Sci. U.S.A.* **75**, 50–53 (1978).
14. Ryan, M. J., Belagaje, R., Brown, E. L., Fritz, H.-J. & Khorana, H. G. *J. molec. Biol.* **254**, 10803–10810 (1979).

15. Prentki, P. & Krishn, H. M. *Gene* **29**, 303–313 (1984).
16. Pouwels, P. H., Enger-Valk, B. E. & Brammar, W. J. *Cloning vectors* (Elsevier, Amsterdam, 1986).
17. Raftery, L. A. & Yarus, M. *EMBO J.* **6**, 1499–1506 (1987).
18. Low, B., Gates, F., Goldstein, T. & Söll, D. *J. Bact.* **108**, 742–750 (1971).
19. Hatfield, D., Diamond, A. & Dudock, B. *Proc. natn. Acad. Sci. U.S.A.* **79**, 6215–6219 (1982).
20. Hatfield, D. *Trends biochem. Sci.* **10**, 201–204 (1985).
21. Mizutani, T. & Tachibana, Y. *FEBS Lett.* **207**, 162–166 (1986).
22. Sunde, R. A. & Evenson, J. K. *J. biol. Chem.* **262**, 933–937 (1987).
23. Buckel, P., Piepersberg, W. & Böck, A. *Molec. gen. Genet.* **149**, 51–61 (1976).

Structure of the liganded T state of haemoglobin identifies the origin of cooperative oxygen binding

Robert Liddington*, Zygmunt Derewenda, Guy Dodson & Deborah Harris

Department of Chemistry, University of York, York YO1 5DD, UK

A molecular description of haemoglobin's cooperative oxygen binding and release was founded on the X-ray crystal structures of the deoxy-T and oxy-R states^{1–3}. Since the R state's oxygen affinity is close to that of an isolated subunit, the crucial allosteric phenomena are (1) the reduced affinity of the T state and (2) the kinetic pathway between the two quaternary structures. To investigate these phenomena directly, we have determined at high resolution ($d_{\min} = 2.1 \text{ \AA}$) the crystal structures of two liganded T-state haemoglobins. In the liganded T-state α subunit, both the tight packing of the haem and the intersubunit contacts inhibit a conformational change between the F helix and FG corner which would allow the haem to become planar and the iron to assume symmetrical R-like coordination. In the β subunit, by contrast, we find no strain on the proximal side, but the intersubunit contacts prevent the haem from tilting about an axis parallel to the F helix which would open up the binding site to oxygen. In both subunits, ligand binding in the T state induces structural changes towards the tertiary conformation of the R state.

Crystallographic study of liganded T-state haemoglobin (Hb) has been hindered in the past by the instability and fracturing of T-state deoxy-Hb crystals grown from high salt following ligation^{4,5}. Relatively stable crystals of partially and fully liganded T-state Hb crystals can, however, be obtained when grown from poly(ethyleneglycol) (PEG)^{6,7}, the form originally studied by Ward and colleagues⁸. Consequently, we have been able to determine at high resolution ($d_{\min} = 2.1 \text{ \AA}$) the isomorphous crystal structures of T-state deoxy Hb and two liganded T-state haemoglobins: T(α -oxy, β -deoxy)Hb [T(α -oxy)Hb]⁷ and T(α -met, β -met)Hb [T(met)Hb]. The structure of the haem environments is essentially identical in T(deoxy)Hb grown from high salt⁹ and in T(deoxy)Hb grown from PEG. The structural changes on binding ligands to the crystals grown from PEG are small compared with the T $\rightarrow$ R transition. Thus, the $\alpha_1\beta_2$ contacts change very little, and the T-state salt bridges remain intact. We shall confine our description to the immediate environment of the haem, to what Gelin and Karplus¹⁰ have called the 'allosteric core'. We compare the structures in two ways, first by overlapping the haem atoms, and second by overlapping the F helix. Table 1 lists the important contact and bonding distances and angles at the haems.

The α subunit. In Fig. 1a and b we see that, relative to the haem, there are insignificant movements in the distal residues and that the α -haem fold remains unaffected in T(α -oxy) and T(met)Hb. The O_2 molecule, which occupies a similar position in both quaternary states, makes essentially identical contacts

* Present address: Department of Biochemistry and Molecular Biology, 7 Divinity Avenue, Harvard University, Massachusetts 02138, USA.

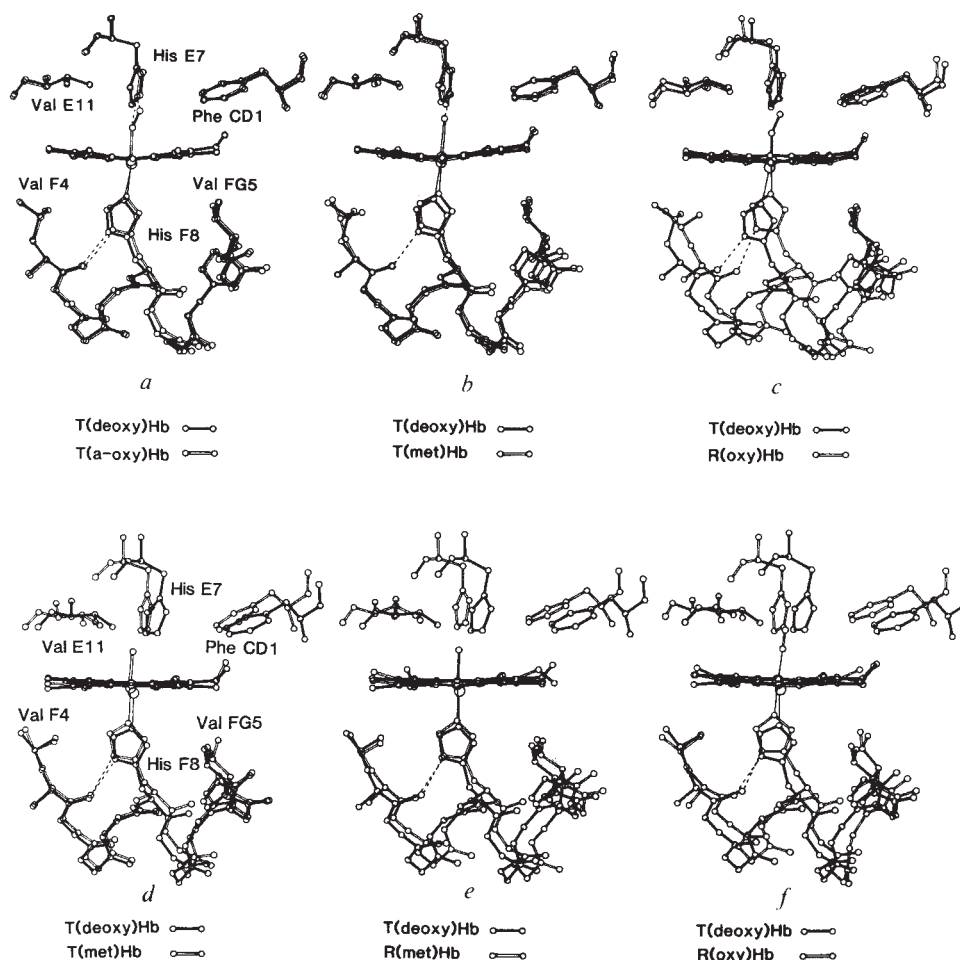


Fig. 1 The haem and surrounding globin structure from the α_1 (a-c) and β_2 (d-f), Hb subunits overlapped onto T deoxy Hb. The haems are matched optimally by least-squares minimization; the propionic acids were excluded from the minimization and are not shown.

in T(α -oxy)Hb and R(oxy)Hb. At the proximal side, we find that the distance of the histidine F8 imidazole ring from the haem plane decreases with increasing levels of ligation, but that any steric interactions in T(met)Hb are essentially the same as those found in R(oxy)Hb⁹: C(E) from N(A) is 2.91 Å in the T state and 2.88 Å in the R state. The stereochemically important movement of the proximal histidine, F8, is its progressive shift towards the axial position: 0.3 Å in T(α -oxy)Hb and 0.5 Å in the T(met)Hb. The Fe atom is also seen to move progressively into the haem plane as the level of ligation increases from T(α -oxy)Hb to T(met)Hb. But the fold in the haem means that although the Fe can move close to the plane of the pyrrole nitrogens it remains out of the overall haem plane. Together these adjustments following ligation alter the coordination geometry towards the symmetrical octahedral structure of the R state (see Fig. 1c). The movements are, however, resisted by the hydrogen bond between F8ND and the main chain carbonyl O at F4.

Figure 2a-c shows the structural changes in the α haem following ligation relative to the F helix, illustrating the important conformational change between the F helix and the FG corner which occurs in the T $\rightarrow$ R transition (Fig. 2c). The movement of the FG corner allows the haem to unfold, providing a simple mechanism for achieving symmetrical octahedral coordination about the Fe atom without disrupting the F helix. In the T state (Fig. 2a, b), the unfolding of the α haem is restrained by the FG corner which makes close contacts with the haem. These restraints can only be released by the quaternary structure change. In T(α -oxy)Hb, movement of the FG corner is barely

significant; in T(met)Hb the movement is about 1/3 that in the T-R transition.

The β subunit. In the T(met)Hb β haem the Fe is oxidized and bound to a water molecule. The change in coordination produces little change in the haem structure (Table 1, Figs 1d and 2d). There is a movement of ~ 0.3 Å by the Fe and the proximal residues towards the haem plane, indicating that the packing of the globin around the β haem is looser than in the α haem. Owing to the more symmetrical character of the fivefold coordinate Fe in the T(deoxy)Hb β subunit, symmetrical sixfold coordination is achieved in T(met)Hb simply by the movement which brings the Fe into the plane of the pyrrole nitrogens to the same extent as in R(met)Hb (B. Luisi, personal communication) Fig. 1d, e. The E helix residues valine E11 and histidine E7, which block the binding site in the T-state deoxy β haem, move markedly with the arrival of ligand in the T state (Figs 1d and 2d). The movement of the E helix still leaves the valine close to the water molecule and the distal histidine, E7, at H-bonding distance.

In the T $\rightarrow$ R transition (Figs 1e, f) the movements at the β haem and the surrounding globin are generally an amplification of the changes produced by ligation in the T state. But a combination of movements is required to clear Val E11 and His E7 from the ligand-binding site when the oxygen molecule arrives: a translation of the E helix across the haem and a tilt of the haem about an axis parallel to the F helix. The tilt is largest in R(oxy)Hb, reduced in R(met)Hb owing to the smaller ligand, and absent in T(met)Hb.

What then do these structural changes in the crystal mean for

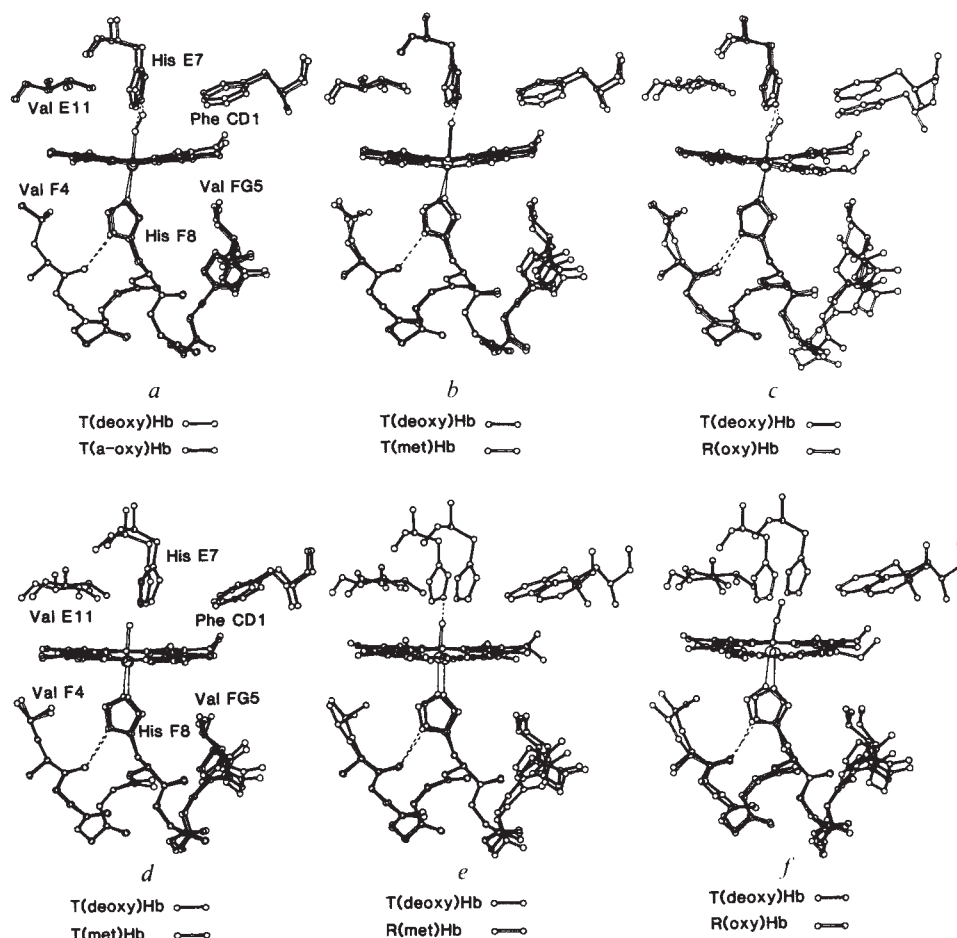


Fig. 2 The haem and surrounding globin of the liganded α and β subunits overlapped onto the T deoxy Hb. The structures are compared using a least-squares fit on the F helix main-chain atoms (F4–F8). *a–c*, the α_1 haem. *d–f*, the β_2 haem.

Table 1 Haem and ligand geometry

				Fe distance (Å) to:								
	Pn	P _{haem}	Pn–Pc	Fe–Ne2		F ₄ O=O–NHF8		Fe–O ₁	O ₁ –O ₂	Fe–O ₁ –O ₂	E ₇ –O ₁	E ₇ –O ₂
T deoxy	0.37	0.57	0.18	α_1	α_2	α_1	α_2					
T α oxy	0.18	0.37	0.18	2.21	2.21	2.55	2.66					
T met	0.09	0.30	0.15(14)*	2.18	2.32	2.83	2.88	1.8†	1.25†	153†	2.8†	2.9†
R met	0.07	0.12	0.05	2.15	2.37	3.06	2.90	2.0†			2.7†	
R oxy	0.12	0.16	0.04		2.04		2.64	2.0				
				1.94		2.75		1.7	1.25	152	2.6	3.0
β subunits												
T deoxy	0.43	0.49	0.05									
	(0.24)*	(0.16)*										
T met	0.09	0.26	0.15	2.17†		2.81†						
R met	0.05	0.17	0.08	2.17†		2.67†		2.0†			2.7†	
R oxy	–0.11	0.00	0.06	2.15		2.68		2.0			3.0	
				2.07		2.75		1.9	1.25	159	3.2	3.5

Pn, plane through pyrrole nitrogens. P_{haem}, plane through haem including first carbon of sidechain. Pc, plane through pyrrole carbons. The data for the three T-state crystals were collected using synchrotron radiation at the Laboratoire pour l'Utilisation de Rayonnement Électromagnétique in Paris and at the Synchrotron Radiation Source in Daresbury (UK). These crystals are isomorphous within experimental error ($P2_12_12$, $a = 95.8$, $b = 97.8$, $c = 65.5 \pm 0.3$ Å). The crystals contain a complete tetramer in the asymmetric unit, providing two independent views of the α and β subunits. Their structures were determined by molecular replacement, using the 2.5 Å structure of Fermi *et al.*¹³ as a model, and refined using several restrained Fast Fourier least squares and difference Fourier methods^{14–17}, to give agreement values ($R = \sum \|F_o| - |F_c| \| / |F_o|$) of between 0.20–0.21 with a root mean square (RMS) deviation from ideal bond lengths of 0.03 Å. In the structure which we denote as T(α -oxy)Hb the α subunits have an oxygen molecule bound to the haem Fe(II) ion; the β haems show small peaks of electron density which may be due to a low level of oxidation. This form was produced by mounting the T(deoxy)Hb crystals in air, after immersion in 33% w/w PEG. The met-haemoglobin crystals-(T(met)Hb) were produced by ferricyanide oxidation of T(deoxy)Hb crystals. In these crystals water molecules are bound to the α and β Fe(III) haems. In T(deoxy)Hb the Fe(II) haems are free from ligand. Judging by crystallographic criteria and structural comparisons, the positional errors are in the range 0.1–0.2 Å for the well-defined protein atoms and ~0.05 Å for the iron atoms.

* The difference between the crystallographically independent α_1 , α_2 , β_1 , and β_2 subunits. These are shown only when they exceed the estimated error (0.15 Å).

† Average values.

the low affinity of T-state Hb and the T→R transition? First, we note that, although modified to some degree by crystal contacts, these changes occur in each of the two crystallographically-independent α and β subunits and thus are likely to be functionally significant. Second, the structural changes are different in the α and β subunits, and so the basis of low affinity is likely to be different. The α haem is packed tightly into its pocket, although there is room for oxygen to bind without steric hindrance. This tight packing prevents the haem from becoming planar and the Fe(II) from assuming R-like symmetrical coordination on binding oxygen. The strain in the ligated α haem is distributed between a nonplanar porphyrin and the H bonds of the F helix and FG corner. In the β subunit, the looser packing of the haem allows the Fe to move into the plane with very little strain on the proximal side. The structure of the haem and its environment in the β subunit of T(met)Hb is intermediate between T(deoxy) and R(oxy), except that there is no tilt of the haem about an axis parallel to the F helix. An inability to tilt when oxygen binds would generate steric strain. Third, although binding of the ligand in the T state is resisted by both haem and globin restraints, it induces changes toward the R-state tertiary conformation in both α and β subunits, as proposed by Gelin *et al.*¹⁸. Thus ligand binding in the T state drives the molecule along the kinetic pathway to the R state. Also the changes that occur in the α subunit in T(met)Hb (where both the α and β haems are liganded) are larger than in T(α -oxy)Hb. This suggests a haem:haem interaction exists within the T state where the loading of the β subunits amplifies the structural responses to ligation in the α subunit.

Our structural findings, particularly in the α subunit, provide direct evidence for Perutz's original hypothesis that the displacement of the iron atom from the haem plane is associated with low affinity¹. We do not know whether the strain we observe in the liganded T state can account fully for the difference in free energy for binding oxygen to the T and R states (~ 3.5 kcal per mole of subunits). For example, theoretical calculations by Warshel and Weiss¹² suggested an upper value of 2 kcal mol⁻¹ for the strain contribution and that the remainder is electrostatic in origin. Accurate haem stereochemistry for the T-state liganded species will provide a better basis for further structural and theoretical research into haemoglobin's cooperative mechanism.

We thank Drs Max Perutz, Ben Luisi, Kiyoshi Nagai, Joyce Baldwin, Cyrus Chothia and Judd Fermi for stimulating discussions, and Dr Rod Hubbard and Mrs Eleanor Dodson for help with computing and graphics. We thank Dr B Luisi for the R(met)Hb refined coordinates, Dr John Helliwell and his colleagues at the SRS (Daresbury) and Drs Roger Fourme and Yves Mauguén at LURE for their assistance in data collection. We thank the MRC and the SERC for support.

Received 7 September 1987; accepted 15 January 1988.

- Perutz, M. F. *Nature* **228**, 726-739 (1970).
- Monod, J., Wyman, J. & Changeux, J.-P. *J. molec. Biol.* **12**, 88-118 (1965).
- Baldwin, J. & Chothia, C. *J. molec. Biol.* **129**, 175-195 (1979).
- Haurowitz, F. Z. *physiol. Chem.* **254**, 266-274 (1938).
- Anderson, L. *J. molec. Biol.* **79**, 495-502 (1973).
- Derewenda, Z. *et al.* in *Proceedings of the Haemoglobin Symposium Faculté des Sciences Université de Bruxelles* (ed. Schneek A. G. & Paul, C.) 52-67 (Edition de l'Université de Bruxelles, Bruxelles, 1983).
- Brzozowski, A. *et al.* *Nature* **307**, 74-76 (1984).
- Ward, L. B., Wishner, B. C., Lattman, E. E. & Love, N. E. *J. molec. Biol.* **98**, 161-171 (1975).
- Fermi, G., Perutz, M. F., Shaanan, B. & Fourme, R. *J. molec. Biol.* **175**, 169-171 (1984).
- Gelin, B. R. & Karplus, M. *Proc. natn. Acad. Sci. U.S.A.* **74**, 801-805 (1977).
- Shaanan, B. *J. molec. Biol.* **171**, 31-51 (1983).
- Warshel, A. & Weiss, R. *J. Am. chem. Soc.* **103**, 446-451 (1981).
- Fermi, G. *J. molec. Biol.* **97**, 225-256 (1975).
- Agarwal, R. C., Lifchitz, A. & Dodson, E. J. in *Refinement of Protein Structures; 'Proceedings of the Daresbury Study Weekend' November 1980* (eds Campbell, J. & Machin, P.) 36-39 (Science and Engineering Research Council, Daresbury, 1980).
- Derewenda, Z. S., Dodson, E. J., Dodson, G. G. & Brzozowski, A. M. *Acta. crystallogr.* **A37**, 407-413 (1981).
- Jack, A. & Levitt, M. *Acta crystallogr.* **A34**, 344-350 (1980).
- Konnert, J. H. & Hendrickson, W. in *Computing in Crystallography* (eds Diamond, R., Ramaschan, S. & Venkatesan, K. 13.01-13.23 (Ind. Inst. Sci., Bangalore, 1980).
- Gelin, B. R., Lee, A. W.-M. & Karplus, M. *J. molec. Biol.* **171**, 489-559 (1983).

Copies of articles from this publication are now available from the UMI Article Clearinghouse.

For more information about the Clearinghouse, please fill out and mail back the coupon below.

UMI Article Clearinghouse

Yes! I would like to know more about UMI Article Clearinghouse. I am interested in electronic ordering through the following system(s):

- ☐ DIALOG/Dialorder ☐ ITT Dialcom
☐ OnTyme ☐ OCLC ILL Subsystem
☐ Other (please specify) _____
☐ I am interested in sending my order by mail.
☐ Please send me your current catalog and user instructions for the system(s) I checked above.

Name _____

Title _____

Institution/Company _____

Department _____

Address _____

City _____ State _____ Zip _____

Phone (_____) _____

Mail to: University Microfilms International
 300 North Zeeb Road, Box 91 · Ann Arbor, MI 48106

Fluorometric calcium measurement

J.S. Hall, K.A. Korkidis and D.L. Maskevich

Measuring the fluorescence of dyes specific for calcium and other cations allows the elucidation of cellular mechanisms such as neurotransmission, muscle contraction and cardiovascular regulation.

A MAJOR focal point of biomedical research over the past decade has been the study of the basic role of Ca^{2+} in physiological events such as neurotransmission, muscle contraction, tumour growth and cardiovascular regulation¹. Work in this area has been advanced greatly by the ability to measure Ca^{2+} in living cells using fluorometric techniques that employ fluorogenic dyes or "probes" with known Ca^{2+} affinity together with specialized commercial instruments which allow the optimal use of these probes^{2,3}.

Fluorometry offers distinct advantages over Ca^{2+} measurements based upon radioactive isotopes, micro-electrodes or photoproteins. Radioactive material presents special handling problems, and micro-electrodes cause cellular damage, measure only local responses, and are limited to use in larger cells. Similarly, calcium-specific photoproteins are slow to disperse and must be introduced by injection, which ruptures cell membranes.

Cation analysis using instruments designed for research with fluorescent probes avoids these drawbacks. Fluorometric analysis is the least biologically disruptive of the techniques thus far devised for intracellular Ca^{2+} measurement. Fluorometry also permits much faster tracking of cation concentration changes, even at millisecond sampling rates. The technique is not confined exclusively to Ca^{2+} ; there are ion-specific fluorescent probes for the measurement of pH and other cations in healthy, living cells^{4,5}. A single instrument system may be utilized for the measurement of cations in macro-environments involving cell populations in suspension or for the micro-measurement of cations in individual cells on a microscope stage.

The most accurate fluorescent cation probes have distinctive dual-excitation or dual-emission characteristics: the fluorescence spectra of these probes change markedly depending upon whether they are free or bound to a cation. Experimental conditions determine the two optimum excitation or emission wavelengths for a specific measurement. At each wavelength there will be a pronounced fluorescence signal corresponding to the free or bound state, and a signal of significantly lower intensity for the opposite state. It is this differential fluorescence intensity that establishes the ratio that enables the measurement of the actual concentration of a cation. Ratioing

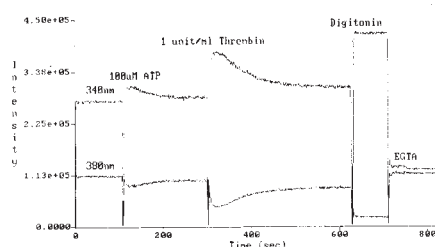


Fig. 1 Fluorescence data acquired at 340 nm and 380 nm from human umbilical vein endothelial cells treated with fura-2.

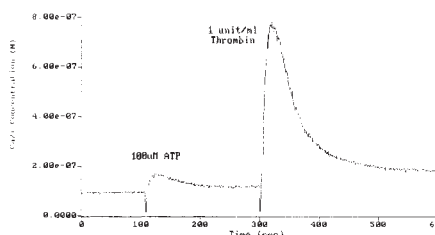


Fig. 2 Changes in the concentration of cytosolic free Ca^{2+} calculated directly from the fluorescence data in Fig. 1.

the two fluorescence signals from a sample also corrects for instrument artefacts, cell-to-cell variations in probe concentration, and reduced fluorescence from photobleaching.

Figures 1 and 2 illustrate the Ca^{2+} flux in human umbilical vein endothelial cells resulting from the sequential addition of the agonists ATP and thrombin. For these measurements, the cells were treated with the dual-excitation probe fura-2. The lysing agent digitonin and the calcium chelator EGTA were used to set the range over which the concentration measurements were made. Figure 1 is a temporal plot of the fluorescence intensity of the sample at excitation wavelengths of 340 and 380 nm. Figure 2 shows the absolute change in the concentration of cytosolic free calcium directly calculated from the data in Fig. 1.

A series of experiments on the contraction mechanism in cultured vascular smooth muscle cells (VSMC) obtained from the aortae of rats by researchers affiliated with Brigham and Women's Hospital and Harvard Medical School demonstrates the sensitivity and flexibility of fluorometric cation measurement^{6,7}. They show that vasoconstrictor agonists induce rapid changes in intracellular Ca^{2+} and pH, and that these ionic events are mimicked by platelet-derived growth factor (PDGF), a potent VSMC mitogen. The fluorescence techniques used in the experiment revealed that Ca^{2+} is mobi-

lized from an intracellular storage site, while a rise in pH is mediated by the activation of Na^+/H^+ exchange.

Although fluorometric cation measurement with dual-excitation probes has become the technique of choice for many researchers, the continuing evolution of hardware and methodology promise even greater experimental utility. Commercial hardware systems are available that incorporate two emission channels for macro- and micro-measurements with dual-emission dyes such as the Ca^{2+} probe indo-1. The recent addition of dedicated digital image processing equipment to existing cation measurement systems gives fluorescence analysis an entirely new dimension by enabling the collection of spatial data about cation migration and concentration.

Without imaging capability, fluorometric analysis of micro-samples can measure only changes in average intracellular fluorescence intensity over time. Image processing can pinpoint the location of site-specific biochemical events within a cell and distinguish the heterogeneous behaviour of cells in an apparently uniform sample. Similarly, fluorescence image processing makes it feasible to monitor ionic changes associated with cell-to-cell communication⁸.

The utility of fluorometric cation measurements is being expanded through the development of a new generation of ion-specific probes. In addition to new specialized calcium dyes, there are probes which promise considerable potential for the measurement of free magnesium, potassium and sodium ions⁹ in many types of living cells, including platelets, renal cells, osteoblasts and myocardial cells. □

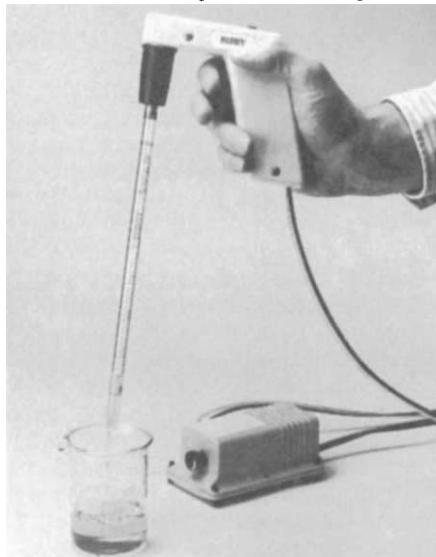
Jonathan Hall, Katherine Korkidis and Dean Maskevich are at Spex Industries, Inc., 3880 Park Avenue, Edison, New Jersey 08820, USA. For more information, fill in reader service number 100.

- Kim, D., Okada, A. & Smith, T.W. *Circulation Res.* **61**, 29 (1987).
- Grynkiewicz, G., Poenie, M. & Tsien, R.Y. *J. biol. Chem.* **260**, 3440-3450 (1985).
- Tsien, R.Y. *Nature* **290**, 527-537 (1981).
- Paradiso, A.M., Tsien, R.Y. & Machen, T.E. *Nature* **325**, 447-450 (1987).
- Paradiso, A.M., Tsien, R.Y. & Machen, T.E. *Proc. natl. Acad. Sci. USA* **81**, 7436-7440 (1985).
- Berk, B.C. *et al. J. biol. Chem.* **262**, 5057-5064 (1987).
- Berk, B.C., Brock, T.A., Gimbrone, Jr., M.A. & Alexander, R.W. *J. biol. Chem.* **262**, 5065-5072 (1987).
- Tsien, R.Y. & Poenie, M. *Trends biochem. Sci.* **11**, 450-455 (1986).
- Minta, A., Harootyan, A.T., Kao, J.P. & Tsien, R.Y. 1987 Annual Meeting of the American Society for Cell Biology, Abstr. No. 492 (1987).

Laboratory measurement notes

A range of instruments for laboratory measurement are described this week, including balances, pipettes, and electrodes for measuring ions and neurotransmitters.

J. BIBBY Science Products Ltd has a **pipetting controller** that turns any pipette into a pipetting gun, speeding repetitive operations (*Reader Service No. 101*). The controller is based on an air pump fitted with a diaphragm system that is operated by a main switch, and provides both positive



Precise pipette control from Bibby.

and negative air flows. It can be used with all pipettes with neck sizes from 2 mm up to 8 mm, including Pyrex glass and Pasteur pipettes. A built-in safety valve prevents fluid from entering the controller head, and the whole unit is fully autoclavable. The Bibby pipetting controller costs £135 (UK).

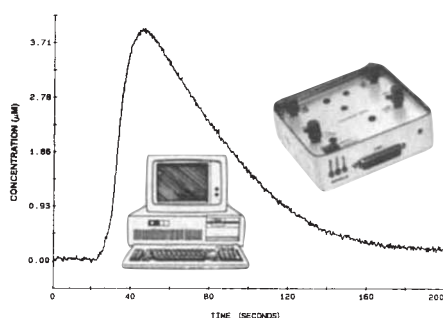
A new line of thirteen **precision balances** is available from the Swiss company Precisa (*Reader Service No. 102*). The different models have a weighing range from 160 g to 12 kg, with reading precisions of 0.001 g, 0.01 g, 0.1 g or 1 g. The balances have a low-profile design, with a height of only 63 mm. A large round weighing pan is supplied with 0.01 g and 0.001 g accuracy models; all other units are equipped with a square platform of 190 × 223 mm. All units in the series have an automatic calibration feature and a capacity/tare ability indicator. Precisa says the balances can be used for parts counting, standard deviation/meanvalue, animal weighing, percentage determination, and converting grams to carats, pounds or penny-weights. Most of the balances in the series have bi-directional digital and analog outputs to allow them to communicate with computers, printers or calculators.

Orion Research Inc. has a new ROSS

sodium electrode, the model 84-11, with a patented internal redox system (*Reader Service No. 103*). Orion says the ROSS electrode is virtually unaffected by temperature variation. In a test of its efficiency, Orion reports that the ROSS electrode reached temperature equilibrium in less than one minute when cycled between two portions of a 100 p.p.m. sodium standard at 25 °C and 80 °C, compared to the three to five minutes required by a conventional sodium electrode to yield a reproducible potential. The ROSS sodium electrode has a reproducibility of + or - 2 per cent standard deviation in successive repetitions, according to Orion. Orion sells the ROSS sodium electrode for \$295 (US).

For performing calorimetric tests including oxidative stability, specific heat, melting behaviour, polymorphism and the determination of glass transition, Mettler Instrument Corp. offers its DCS30S **differential scanning calorimetry module** and furnace (*Reader Service No. 104*). Mettler says the conductive properties of the module's silver furnace gives greater baseline reproducibility and a more rapid transition to temperature equilibrium. A liquid nitrogen cooling apparatus allows sample loading at very low temperatures and continuous measurement without switching modules, through a temperature range of -170 °C to +600 °C. Mettler sells the differential scanning calorimetry module for \$15,150 (US).

An *in vivo* electrochemistry system is available from Medical Systems Corp. for the measurement of **extracellular neurochemical changes** in the brain at speeds



The *in vivo* electrochemistry system from Medical Systems Corp. measures neurotransmitters in living brain tissue.

approaching physiologic events (*Reader Service No. 105*). The \$6,995-13,500 (US) model IVEC-5 system sheds light on the dynamics of neural mechanisms by recording the concentration of neuro-

transmitters such as dopamine, serotonin and noradrenaline at speeds of up to ten times per second, says Medical Systems Corp. The instrument measures electrical currents through microelectrodes placed at the prepared site, amplifies the signal, and transmits it to a computer where the system software analyses and plots the neurochemical concentrations in real time.

For verifying the accuracy and precision of pipetting and diluting equipment, Cavo offers a **dilutor performance check kit** based on a dye dilution technique (*Reader Service No. 106*). Cavo says the check kit can determine the accuracy of any dilution ratio from 1:4 to 1:500, with results obtained using a standard curve prepared from precise standards provided with the kit. To perform the procedure, a dye concentrate is diluted with a dilution buffer with the pipetter to be checked. The absorbance of the diluted dye is read in a spectrophotometer and compared with a standard curve. The accuracy of the dilution is then determined by comparing the absorbance obtained with a standard curve. The \$50 (US) check kit contains a matched set of reagents sufficient to develop four standard curves and a minimum of 60 dilution tests.

The model 670 **titroprocessor** from Metrohm allows the user to program a titration sequence exactly, with video screen dialogue (*Reader Service No. 107*). Four titrating burettes and 12 dispensing burettes are controlled from the instrument, allowing complex titrations of up to four different titration procedures to be combined into a single method. Both dynamic and monotonic titrations having up to nine inflection points can be executed, in addition to titrations to fixed endpoint values and moisture determinations according to Karl Fisher. Data from failed titrations can be reprocessed after parameters controlling the recognition of endpoints are adjusted, until the inflection point or points have been identified. Data displayed on the screen can be stored or printed out on the built-in A4 thermal printer/plotter. Interfaces for the connection of balances, computers and a range of sample changers are included. □

These notes are compiled by Carol Ezzell from information provided by the manufacturers. To obtain further details about these products, use the reader service card bound inside the journal. Prices quoted are sometimes nominal and apply only within the country indicated.